

Integrating distributed post-genomic data to infer the molecular basis of bacterial phenotypes

Thesis by

Tracy Craddock

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

NEWCASTLE UNIVERSITY LIBRARY

206 534 19 4

Thesis L8706



Newcastle upon Tyne, UK

2008

(Defended December 7, 2007)

To Mum, Dad and Steve.
And Peter.

Abstract

The aim of the project described in this thesis is to understand and predict the characteristics and behaviour of a family of bacteria through an analysis of genome wide data from a variety of sources.

The focus of the research is a family of bacteria, *Bacillus*, whose members show a diverse range of phenotypes, from the non-pathogenic *B. subtilis* to *B. anthracis*, the causative agent of anthrax. Specifically, the focus was on the genomic scale identification and characterisation of secreted proteins from *Bacillus* species.

Firstly, the application of Grid-based computational approaches to problems in genomic analysis and annotation was investigated, applying m^2 Grid technology to a biological problem not previously addressed using this approach. e-Science workflows and a service-oriented approach were developed and applied to predict and characterise secreted proteins, and the results automatically integrated into a custom relational database. An associated Web portal was also developed to facilitate expert curation, results browsing and querying over the database. Workflow technology was also used to classify the putative secreted proteins into families and to study the relationships between and within these families. The design of the workflows, the architecture and the reasoning behind the approach used to build this system, called BaSPP, are discussed.

Analysis of the putative *Bacillus* secretomes revealed clear distinctions between proteins present in the pathogens and those in the non-pathogens. The properties of the protein families present in all *Bacillus* secretomes, as well as those specific either to the pathogens or to the non-pathogens were investigated.

Many of the protein families contained members of unknown function. In the

second part of the project, these families were investigated in more depth, using additional data integration strategies not previously applied to these organisms. The secretomes were modelled at the system level, in the broader context of interactomes. A system called SubtilNet was therefore developed, using *B. subtilis* as the reference organism. As part of SubtilNet, a toolkit and architecture were developed and implemented for building and analysing probabilistic functional integrated networks (PFINs). The PFINs built for each *Bacillus* species using this system were subsequently used to delve further into the interactions specific to the secreted proteins by extracting and exploring the cross-species PFINs of these proteins. The cross-species PFINs for the protein families specific to the pathogens and non-pathogens were explored, with particular emphasis on the core PrsA-like protein family, which acted as a use case to show how the PFINs can be used to shed light on protein function. The addition of orthologous links between species was demonstrated to facilitate network clustering and analysis, enabling putative annotations to be applied to proteins previously of unknown function.

Declaration

I hereby declare that this thesis, apart from the help recognised, is my own work and has not been formerly submitted to another University for a degree.

Tracy Craddock

September 2007

Acknowledgements

This thesis would not have been possible without the support of a number of people. In particular, I would like to extend a special thanks to my supervisor Dr Anil Wipat, without whom this would not have been possible. Special thanks also go to Prof Colin Harwood for his expert guidance on the biological aspects of this thesis. I am also grateful to Dr Robert Hirt for his biological support. Thanks also to the help and advice of Dr Phillip Lord and Dr Jennifer Hallinan. For their help and assistance, I would also like to thank Dr Matthew Pocock and Allyson Lister. For their technical support, I would like to extend a thanks to Dr Daniel Swan and Keith Flanagan. Last, but by no means least, I would also like to thank my office buddies Frank Gibson, Dr Olly Shaw and Keith Hayward for making it an enjoyable experience.

I also gratefully acknowledge the support of the North East Regional e-Science Centre and the European Commission (LSHC-CT-2004-503468). I also thank the EPSRC and Non-Linear Dynamics for their support.

Contents

Abstract	iii
Declaration	v
Acknowledgements	vi
Abbreviations	xii
1 Introduction	1
1.1 Computational challenges in systems biology	2
1.2 e-Science and Grid technology	3
1.3 High-throughput genomic characterisation	5
1.3.1 <i>my</i> Grid and e-science workflows	6
1.4 Secretome analysis of bacteria	8
1.5 Elucidating protein function using integrated networks	9
1.6 Aims and objectives	10
1.6.1 Aims	10
1.6.2 Objectives	11
2 Background	13
2.1 Introduction	13
2.2 <i>Bacillus</i> genomes and their secretome	13
2.2.1 Typical bacteria	13
2.2.2 Gram-positive vs. Gram-negative bacteria	14
2.2.3 <i>Bacillus</i> species	15

2.2.4	Mechanisms of secretion	20
2.2.4.1	Amino-terminal signal peptides	22
2.2.4.2	Translocation machinery	25
2.2.4.3	Retention mechanisms	27
2.3	Data integration, e-science and Grid technology	30
2.3.1	e-Science and the Grid	32
2.3.1.1	Architecture	34
2.3.2	Services on the Web	37
2.3.2.1	Soap-based Web services	37
2.3.2.2	REST services	40
2.3.3	Taverna and <i>m</i> yGrid	41
2.3.4	Microbase	43
2.4	Biological network analysis	44
2.4.1	Interactome modelling	45
2.4.2	Interactome comparison	47
2.4.3	Interactome visualisation	50
2.5	Conclusions	52
3	An e-science approach to the genomic scale characterisation of bacterial secreted proteins	54
3.1	Introduction	54
3.1.1	Computational approaches to predicting and classifying secreted proteins	55
3.2	Results	56
3.2.1	System architecture	56
3.2.2	Development of a secretory protein classification workflow . .	59
3.2.2.1	Workflow design and implementation	62
3.2.2.2	Comparison of BaSPP-based predictions for <i>B. subtilis</i> secreted protein to experimentally verified protein annotations	68

3.2.2.3	Comparison of BaSPP-based predictions for <i>B. subtilis</i> secreted protein to previously published, computationally generated, annotations	72
3.2.3	Protein analysis workflow	74
3.2.3.1	Workflow design and implementation	76
3.2.3.2	Comparison of <i>Bacillus</i> secretory protein families to COGs	78
3.2.4	Functional mapping of GO terms to the SubtiList classification hierarchy	78
3.3	Discussion	81
4	Properties of the predicted secretomes of members of the genus <i>Bacillus</i>	86
4.1	Introduction	86
4.2	Results	87
4.2.1	Overview of the predicted secretome composition of 12 <i>Bacillus</i> genomes	87
4.2.2	Taxonomic similarity between the <i>Bacillus</i> species based on predicted secretome composition	90
4.2.3	Functional analysis of the predicted secretomes of the <i>Bacillus</i> species	93
4.2.3.1	Core families	94
4.2.3.2	Families specific to the non-pathogen	94
4.2.3.3	Families specific to the known pathogen	98
4.3	Discussion	112
5	The construction of probabilistic functional integrated networks (PFINs) for members of the genus <i>Bacillus</i>	120
5.1	Introduction	120
5.2	Methodology	121
5.2.1	Unified scoring method	121

5.2.2	Gold standard	123
5.2.3	<i>B. subtilis</i> PFIN	123
5.2.3.1	Datasets	123
5.2.3.2	Integration of datasets	128
5.2.4	Other <i>Bacillus</i> species PFINS	128
5.2.5	Identification of clusters within <i>Bacillus</i> PFINS	129
5.3	Results	131
5.3.1	Threshold determination for network optimisation	131
5.3.2	Network properties of <i>Bacillus</i> PFINS	131
5.3.3	Determination of optimal clustering settings of the <i>B. subtilis</i> PFIN	142
5.3.4	Functional analysis of the clustered <i>B. subtilis</i> PFIN	143
5.3.5	Distribution of secreted proteins within the topology of the <i>B.</i> <i>subtilis</i> PFIN	143
5.4	Discussion	147
6	The application of PFINS to the systems level analysis of secreted protein families	149
6.1	Introduction	149
6.2	Methodology	149
6.3	Results	150
6.3.1	Analysis of cross-species clusters incorporating protein families	150
6.3.2	A detailed cross-species analysis of the core PrsA protein family and interacting partners using PFINS	159
6.3.2.1	The PFIN for the core PrsA protein family	159
6.3.2.2	The global structure of PFIN for the core PrsA protein family and functional interacting partners	162
6.3.2.3	Cluster analysis of the PrsA protein family PFIN	164
6.4	Discussion	167

7	General discussion, conclusions and future work	173
7.1	Discussion	174
7.2	Conclusions	179
7.3	Future Work	180
A	Putative <i>B. subtilis</i> secreted proteins	216
B	SubtiList classification codes	227
C	SubtilNet GO level threshold deduction based on <i>B. subtilis</i>	230
D	SubtilNet log likelihood calculations for <i>B. subtilis</i>	231
E	SubtilNet network topologies	235
F	<i>B. subtilis</i> PFIN clusters	247
G	Cross-species PFINs	268

Abbreviations

API Application Programming Interface

ban *B. anthracis* (Ames, isolate Porton)

bar *B. anthracis* (Ames ancestor)

bat *B. anthracis* (Sterne)

bce *B. cereus* (ATCC 14579/DSM 31)

bcl *B. clausii* (KSM-K16)

bcz *B. cereus* (ZK/E33L)

bli *B. licheniformis* (DSM 13/ATCC 14580)

bsu *B. subtilis* (strain 168)

btk *B. thuringiensis konkukian* (strain 97-27)

BaSPP Bacterial Secretory Protein Prediction

BPEL4WS Business Process Execution Language For Web Services

COGs Clusters of Orthologous Groups of proteins

DAG Directed Acyclic Graph

DR Domain Enrichment Ratio

GO Gene Ontology

GOA GO Annotations

GUI Graphical User Interface

HMM Hidden Markov Model

HTTP HyperText Transport Protocol

IETF Internet Engineering Task Force

KEGG Kyoto Encyclopedia of Genes and Genomes

MCL Markov Clustering

MCODE Molecular Complex Detection

OGSA Open Grid Services Architecture

PFIN Probabilistic Functional Integrated Network

PTS Phosphotransferase System

RBH Reciprocal Best Hit

RDP Ribosomal Database Project

REST Representational State Transfer

RIO Resampled Inference of Orthology

RNSC Restricted Neighbourhood Search Clustering

RPC Remote Procedure Call

RSD Reciprocal Smallest Distance

SCUFL Simple Conceptual Unified Flow Language

SPase Signal Peptidase

SPPases Signal Peptide Peptidases

SpI Type I Signal Peptide

SpII Type II Signal Peptide

SPC Super Paramagnetic Clustering

UDDI Universal Description, Discovery, and Integration

URI Universal Resource Identifier

URL Uniform Resource Locator

W3C World Wide Web Consortium

WSDL Web Service Description Language

XML eXtended Markup Language

Chapter 1

Introduction

One of the biggest challenges for biology is to understand the complexity of biological systems in terms of the interaction of their components, the functions that these interactions give rise to, and the consequent behaviour of the system [Nurse, 2003]. Within genomics, a great deal of information can be obtained about the biology of an organism by analysing its genome, and subsequently interpreting this analytical data in the context of existing genome sequence information [Gabaldón and Huynen, 2004]. However, genomics cannot yet provide us with a complete understanding of many aspects of biological function, as it is difficult to predict the behaviour of gene products from gene sequences alone [Dove, 1999].

New post-genomic approaches are therefore required to bridge the gap between the genome and cellular behaviour [Dove, 1999]. Knowledge at a variety of different levels is required that provide data on many types of cellular components, including the transcriptome (complete set of transcripts), proteome (complete set of proteins), interactome (complete set of interactions) and localisome (localisation of all transcripts and proteins) [Ge et al., 2003]. Integrating the information extracted from these additional omics approaches provides a bridge to a discipline of biology known as *systems biology*. By understanding the complex interactions of genes, proteins, and so forth, knowledge of a biological system (e.g. metabolic pathways) is obtained, leading to more precise functional annotations for gene products and relationships [Kitano, 2002]. In this study, the application of this approach to integrative biology, in combination with an e-science approach, is demonstrated through investigation of

the secretome of a genus of Gram-positive bacteria, *Bacillus*.

1.1 Computational challenges in systems biology

With the abundance of information available on the inner workings of the cell, the challenge is how to extract biological meaning about systems as a whole from these multiple omic datasets. Systems biology addresses these biological questions by integrating omic datasets to create computable mathematical models representing interesting biological phenomena. These models define a set of assumptions and hypotheses that need to be tested or confirmed experimentally. Systems biology is therefore both a data- and hypothesis-driven science [Aderem, 2005, Joyce and Palsson, 2006].

Computational "dry" experiments (such as simulation) on the biological models may agree with the embedded assumptions and hypotheses of the model, or otherwise highlight inconsistencies with experimentally verified knowledge. Inconsistent models may result in model rejection or modification. Consistent models may lead to the formation of a number of predictions that can then be used to drive "wet" experiments. Analysis of the results from the wet-lab experiments can subsequently verify the models predictions [Aderem, 2005, Kitano, 2002].

Systems biology therefore combines omic approaches, data integration, modelling and wet-lab biology [Ge et al., 2003]. Developments in this area herald a new era in which the benefits of comparative analysis of the rapidly growing collection of complete genomes will complement experimental approaches aimed at improving our understanding of biological systems [Subramanian et al., 2001]. The development of scalable computing systems that can rapidly analyse and compare sequenced genomes in an efficient manner is required to assist in the knowledge gathering process. The development of high-throughput technologies and the desire to understand organisms as complex systems, rather than the more traditional approach of studying their component parts, is placing new requirements on computing infrastructure. Managing the conduct of high-throughput experiments, the data generated by them, and the sharing and integration of these datasets enables better-informed or faster decisions.

Computational methods can therefore guide and focus traditional wet-lab biological research, as computational analysis provides interesting insights into genomic differences [Aderem, 2005, Subramanian et al., 2001]. The use of computational methods for analysis falls into an area of science known as *bioinformatics*. In general, the term bioinformatics means the application of information technology to the management and analysis of biological data [Attwood and Parry-Smith, 1999]. It is therefore a multidisciplinary research area at the interface between informatics and biology. Bioinformatics therefore plays an important role in systems biology.

1.2 e-Science and Grid technology

Within bioinformatics, a considerable amount of time is spent dealing with three problems: heterogeneity, distribution and autonomy. Bioinformatics tools and datasets tend to be highly heterogeneous in content and structure; the datasets are often widely geographically dispersed, and are often autonomously controlled as they are maintained and deployed by individual scientists within their own laboratories [Stevens et al., 2007]. e-Science technologies promise to help to address these issues.

”e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it” [Taylor, <http://www.nesc.ac.uk>].

e-Science therefore refers to a science in which researchers from around the world collaboratively share distributed resources. The use of an e-science approach enables faster, better-informed research by providing access to resources located on distributed computers as if they were on the user’s own machine. The resources can include enormous computing power used for large calculations, as well as databases and the necessary software to allow this information to be shared and integrated, encouraging collaboration amongst researchers from different institutions. Through e-science it becomes possible to conduct research that was previously impossible using a single computer. Bioinformatics is therefore a discipline for which e-science is appropriate [Hey and Trefethen, 2005].

The infrastructure supporting e-science is termed the *Grid* [Hey and Trefethen, 2003, 2005]. The Grid is based upon coordinated resource sharing and problem solving in virtual organisations. The Grid tackles many concerns and requirements not addressed by current distributed technologies, such as the need for flexibility and control over how a variety of resources (programs, data, computers, sensors and networks) are shared and used, as well as issues concerning quality of service, scheduling and accounting [Foster et al., 2001]. The Grid provides a means for scientists working on large distributed projects to perform experiments using an array of computational resources in order to solve scientific problems [Greenwood et al., 2003].

It should be noted that the Grid itself is not e-science, because in order for the Grid to support a scientist performing analysis and simulations upon a computer, more layers of software are needed on top of the basic computer technology that forms the Grid. This software, termed *middleware*, lies conceptually between the operating system software and the applications software that is designed to solve a particular problem for a user. The function of the middleware is to organise and integrate distributed and heterogeneous Grid resources to provide a "level-playing field" for e-scientists to run applications on suitable distributed computers [CERN, <http://gridcafe.web.cern.ch>].

Middleware development is the focus of many ongoing Grid projects worldwide, resulting in a number of prototype systems for bioinformatics problems, such as *myGrid*¹. *myGrid* is an e-science service based middleware toolkit used to build Grid enabled bioinformatics applications. *myGrid* aims to support biology and its sub-disciplines. It essentially illustrates how Grid technology can be harnessed and enhanced to accommodate the needs of biologists and a wider range of e-scientists than was the original target audience of the Grid [Stevens et al., 2003]. The ultimate goal of this project was to develop open source high-level middleware to support the construction, management and sharing of personalised data-intensive *in-silico* experiments in biology on a Grid [Greenwood et al., 2003, Stevens et al., 2003]. Using *myGrid*, e-scientists can use Web services, which provide a standard way of connecting different software

¹*myGrid* Website: <http://www.mygrid.org.uk/>

applications that are located on different platforms or frameworks, to construct their *in-silico* experiments [W3C, <http://www.w3.org>].

1.3 High-throughput genomic characterisation

Many bioinformatics experiments can be represented as a series of workflows, integrating a number of programs and data sources to test a hypothesis *in-silico*. Workflows are essential to allow e-scientists to capitalise on the increasing number of resources being exposed as Web services, enabling the coordinated use of a variety of distributed and heterogeneous resources. These workflows, alternatively called pipelines, form the bedrock of computational analysis within bioinformatics [Addis et al., 2003, Oinn et al., 2006]. Pipelines are a subset of the more general term workflow, in which there exists added layers of control in the workflow.

Workflows that integrate a number of bioinformatics software tools with multiple databases allow the analysis and storage of genome sequences to be automatically managed. One of their most common applications in bioinformatics is to support and complement the manual annotation process; an example is Ensembl [Curwen et al., 2004]. The primary goal in developing annotation workflows is to provide highly accurate and reliable results using the widest possible range of evidence from available databases. Current annotation workflows utilise the consensus based approach in which a genome sequence passes through several successive levels of analysis, each consisting of a number of different algorithms. By applying this approach, new complex systems can be built from smaller component programs and data sources. The component programs and data sources can in theory be located anywhere in the world [Addis et al., 2003, Rust et al., 2002]. However, there are practical problems limiting the location of databases and resources, as conventional technology requires these to be housed in a single location, due to problems relating to coordination and distribution, but e-science technology and workflows can overcome these limitations.

A possible, and rather simple, structure of an annotation workflow is depicted in figure 1.1. Within this possible architecture, the core database distributes data

between the analysis workflow on the left, and the Website on the right. The Website integrates the annotated data stored in the core database with multiple, external databases. In any workflow, the general starting point is the processing of raw sequence data. Gene locations can subsequently be predicted by aligning homology search matches and making computational predictions on the genomic sequence. Following the identification of predicted genes, the function of the genes needs to be determined. The function of genes can be inferred by mapping predicted gene sequences to known genes and proteins in sequence databases using similarity search algorithms (e.g. BLAST [Altschul et al., 1990]) or by using protein domains (e.g. Interpro [Quevillon et al., 2005]). The final stage is to distribute the annotated information to the users. The most common means of doing this is via a Website [Rust et al., 2002], although other systems exist [BioDAS, <http://www.biodas.org>].

When developing a workflow, a number of factors must be taken into account. There is continuing improvement of the quality of sequences and assemblies, and consequently the replacement of redundant sequences with new, superior sequences. An automated workflow must therefore allow new sequences to be added and analysed without disruption. In addition, a workflow must be extensible, so that new or improved algorithms and methodologies can be inserted into the analysis process without redesign of the system [Addis et al., 2003, Rust et al., 2002].

1.3.1 *my*Grid and e-science workflows

Traditionally, workflows have been implemented in one of two different ways. In those laboratories with the necessary resources or specialist support, they have often been automated using Perl, which is ideally suited to the manipulation of the textual representations that biology has traditionally used to store its data. The majority of biologists, however, have used cut-and-paste between the myriad of Websites offering access to underlying computational resources. e-Science projects such as *my*Grid aim to provide an alternative to these two approaches for integrating programs and data, making use of Web services. The use of Web services, a workflow enactment engine

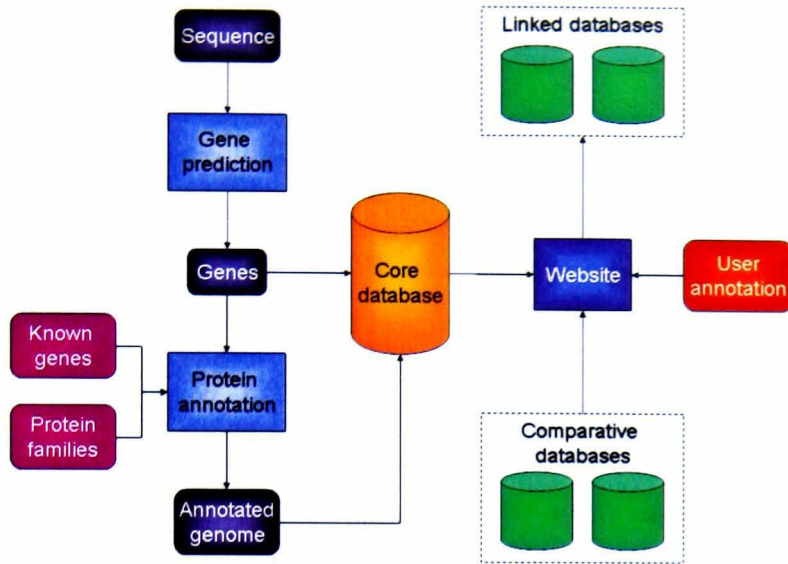


Figure 1.1: The basic structure of an automatic genome annotation workflow. Adapted from Rust et al. [2002].

and a convenient, easy-to-use workflow editor, known as Taverna², have enabled lab biologists to access some of the power of automation previously only available to programmers [Oinn et al., 2006].

myGrid provides the capability to integrate a vast range of resources, in terms of data and applications that may be available within an organisation or as external services, into a workflow. *myGrid* has therefore built services such as resource discovery, workflow enactment and distributed query processing. Additional services, such as provenance management, change notification and personalisation have also been developed. These services are often neglected at the workstation, but are required to support the scientific method and best practice found at the bench [Stevens et al., 2003]. *myGrid* has previously been applied to biological questions, including problems relating to Graves' Disease [Li et al., 2004] and William-Beuren Syndrome [Stevens et al., 2004b].

²Taverna Website: <http://taverna.sourceforge.net/>

1.4 Secretome analysis of bacteria

One of the main mechanisms that bacteria use to interact with their environment is to synthesise proteins and export them from the cytoplasm (the site of synthesis) to an extracytoplasmic location. Secreted proteins are often important in the survival of a bacterium in a particular environment. The entire complement of secreted proteins is often referred to as the *secretome*. Characterising secreted proteins and the mechanisms of their secretion can reveal a great deal about the capabilities of an organism. Soil organisms secrete enzymes to recover nutrients from their immediate surroundings [Tjalsma et al., 2000]. During infection, many pathogenic bacteria secrete virulence proteins (including enzymes and toxins) into their extracellular environment can subvert the host defence systems, for example by inactivation of elements of the innate immune system, and facilitate the entry, movement and dissemination of bacteria within host tissues [Nomura and He, 2005]. Bacteria may also use secreted peptides to communicate with each other, allowing them to form complex communities and to adapt rapidly to changing conditions [Piazza et al., 1999]. The secretomes of bacteria are therefore of great importance and interest.

Recently, our understanding of the secretome has greatly improved due to the availability of increasing numbers of complete bacterial genome sequences, essentially molecular blueprints containing the information that describes the entire protein repertoire of an organism. Armed with these sequences, it is possible to predict which proteins are likely to be secreted, as well as the mechanisms of their secretion.

Biologists have already begun to apply conventional bioinformatics technology to the prediction and classification of secreted proteins. The first, largely genome-based survey of a secretome was carried out using bioinformatics tools on the genome of *Bacillus subtilis* (strain 168) [Tjalsma et al., 2000], using legacy tools called from custom scripts in combination with expert curation. A number of studies have consequently been undertaken focusing on various other organisms, in which bioinformatics programs and resources play a vital role [Binnewies et al., 2005, Boekhorst et al., 2006, Lewenza et al., 2005]. Comparative secretomics can also be used to reveal information

about the core and variant secretomes of bacteria [Sibbald et al., 2006].

These studies involve the repeated application of a number of different algorithms to all of the gene and protein sequences encoded on the genome. Many of these algorithms are computationally expensive and, given that an average bacterial genome can encode around 4,000 or more proteins, the process can become computationally bound. In addition, the results of the application of these algorithms needs to be stored and integrated in order to make a prediction about the secretory status of the entire set of proteins encoded by a particular genome. Often, the results of the classification algorithms may be error prone and mechanisms to permit expert human curation and results browsing also need to be established.

By applying e-science workflows and a service-oriented approach to the genomic scale detection and characterisation of secreted proteins, the e-scientist can describe and launch experimental processes in a structured, repeatable and verifiable way [Addis et al., 2003].

1.5 Elucidating protein function using integrated networks

The ability to elucidate protein function is a common problem in the post-genomic era. High-throughput sequencing analysis is still revealing large numbers of new genes of unknown function. A common approach to computationally assigning function is based on querying databases to identify similar sequences with known function (e.g. BLAST) or by identifying functional domains (e.g. Interpro).

However, alternative approaches are required when a protein shows no significant similarities or identifiable domains. More recently, an additional strategy has been employed, by identifying all functional protein-protein interactions. Originally, these protein-protein interactomes were generated based on one data source. However, interactomes have advanced in recent years towards combining multiple data sources, providing functionally integrated networks [Joyce and Palsson, 2006]. Protein-protein

interactomes can be used to infer the function of uncharacterised proteins in the context of its neighbouring environment [Sun et al., 2006, Vazquez et al., 2003]. Alternatively function can be inferred by making comparisons between multiple interactomes to identify previously unknown orthologues [Chen et al., 2007] or missing functional links.

Studies have been undertaken in which a variety of protein-protein interactomes have been developed for a number of different organisms, incorporating information from various omic data sources [Date and Stoeckert, 2006, Deng et al., 2004, Jansen et al., 2003, Kiemer et al., 2007, Lee et al., 2004, Rhodes et al., 2005, von Mering et al., 2007]. However, solutions are required to better conceptualise and interpret the vast amount of omic information used to construct interactomes. The aim is to take the data from different types of studies and produce detailed models of complex systems, particularly biological pathways. In order for this to occur, the datasets have to go through some complicated analysis and visualisation, which can be CPU-intensive. There is therefore a need for tools that take dissimilar datasets and provide 3D visualisations of complex systems with a high degree of definition and accuracy.

1.6 Aims and objectives

1.6.1 Aims

This project aimed to research aspects of Grid based computational systems designed to integrate and analyse data arising from post-genomic technology. The application of Grid based computational systems to problems in genomic analysis and annotation were investigated.

Through this work, the application of *mv*Grid technology to an additional biological problem than has previously been described was investigated. The aim was to understand and predict the characteristics and behaviour of a family of bacteria, through an analysis of their complete genomic sequences. The focus of this study was on a family of bacteria, *Bacillus*, whose members show a diverse range of properties.

Proteins of unknown function were therefore of particular interest. A means of annotating proteins with a putative function was consequently required. To provide this capability, novel data integration strategies for identifying protein function at a systems level were explored.

1.6.2 Objectives

Specifically, this thesis describes the development and application of e-science workflows and a service-oriented approach to the genomic scale detection and characterisation of secreted proteins from *Bacillus* species. This problem places different requirements on the workflow and the surrounding architecture than has previously been described in the bioinformatics workflow domain.

Firstly, a novel data mining, integration and knowledge derivation annotation workflow was developed to analyse bacterial proteomic data so as to predict and characterise secreted proteins. This workflow harnesses provenance generated through execution of the workflow. The data generated at key steps in the workflow was stored in a database. A Web interface was then developed to provide a means of viewing and curating the data within the database.

Following the classification of proteins, a novel analysis workflow was developed to compare the predicted characteristics and behaviour of *Bacillus* species to make further predictions about their pathogenesis and phenotypes. Both workflows were constructed using the ^{my}Grid workbench Taverna. As with any workflow, these pipelines were designed to be flexible and extensible to allow for future adjustments. The system was utilised to make predictions about the secretomes of the only 12 *Bacillus* isolates for which complete genomic sequences were publicly available from Genome Reviews³ at the time of download^{4 5}; this includes: *B. anthracis* (Sterne), *B. anthracis* (Ames ancestor), *B. anthracis* (Ames, isolate Porton), *B. cereus* (ZK/E33L), *B. cereus* (ATCC 10987), *B. cereus* (ATCC 14579/DSM 31), *B. clausii* (KSM-K16), *B.*

³Genome Reviews Website: <http://www.ebi.ac.uk/GenomeReviews/>

⁴Downloaded: April 2006

⁵In August 2007, there were 14 *Bacillus* isolates with complete genomic sequences

halodurans (C-125/JCM 9153), *B. licheniformis* (DSM 13/ATCC 14580, sub_strain Novozymes), *B. licheniformis* (DSM 13/ATCC 14580, sub_strain Goettingen), *B. subtilis* (strain 168) and *B. thuringiensis konkukian* (strain 97-27).

A set of probabilistic functional integrated networks (PFINs) were then developed for these *Bacillus* species, providing a novel approach to characterise the putative secreted proteins of unknown function, and to make further predictions about their involvement in bacterial pathogenesis and phenotypes.

The remaining chapters of this thesis are therefore divided as follows:

- Chapter 2 provides background and a literature review relating to the important concepts and previous research in the areas of work covered by this thesis.
- Chapter 3 describes the construction of the e-science workflows used to classify and analyse the secreted proteins of Gram-positive bacteria.
- Chapter 4 presents an analysis of the predicted secretomes generated by the application of the workflows to 12 *Bacillus* genomes.
- Chapter 5 describes the development and characterisation of functional integrated networks for 12 *Bacillus* strains.
- Chapter 6 describes the application of functionally integrated network comparison techniques to the study of selected secretory protein families.
- Chapter 7 discusses the findings and conclusions that can be drawn and discusses possible future work.

Chapter 2

Background

2.1 Introduction

The biological questions addressed in this work focus on identifying and characterising the secretome of Gram-positive bacteria, specifically in terms of the *Bacillus* species. The first section of this chapter therefore introduces the *Bacillus* species, along with an insight into the mechanisms used by these bacteria to secrete proteins. The next section discusses e-science and associated technologies most applicable to the development of an e-science based system capable of classifying and analysing the secretomes of Gram-positive bacteria. Finally, strategies for developing, analysing and visualising integrated functional networks are presented.

2.2 *Bacillus* genomes and their secretome

In order to understand the mechanisms involved in protein secretion, an appreciation of the structure of a bacterial cell is required. The distinction between the two different groups of bacteria (Gram-positive and Gram-negative) should also be noted.

2.2.1 Typical bacteria

A typical bacterium is generally composed of a number of different structures and compartments, as depicted in figure 2.1. The *cytoplasm* is the region within the bacterium where the majority of cellular functions take place; the *nucleoid* is the region

within the cytoplasm where the chromosome (usually a single molecule of DNA) is located; *ribosomes* are small particles in the cytoplasm involved in protein synthesis. The *cytoplasmic membrane* is a layer of phospholipids and proteins that act as a selective permeability barrier. The *cell wall* is a rigid structure giving the cell shape, as well as protecting the inner components of the cell from damage. The cell wall may also be surrounded by an outer membrane depending on the type of bacterium. The *capsule* is an outer layer associated with some bacterial species, providing additional protection. The *pili* are hair-like structures on the surface of a bacteria that aid in the attachment to other surfaces. The *flagella* are hair-like structures involved in locomotion [Hale et al., 1995, Sonenshein et al., 2002] [Davidson, <http://micro.magnet.fsu.edu>].

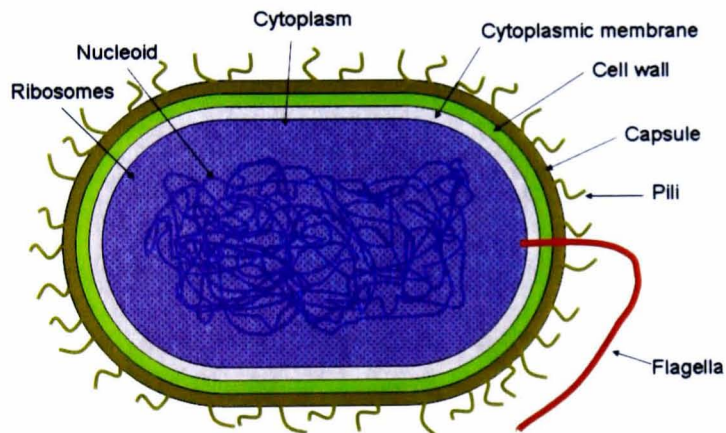


Figure 2.1: The cell structure of a typical bacterium. Adapted from Davidson, <http://micro.magnet.fsu.edu>.

2.2.2 Gram-positive vs. Gram-negative bacteria

Bacteria can be differentiated into two distinct groups (Gram-positive and Gram-negative) based on the chemical and physical properties of their cell walls (figure 2.2). The method employed for this purpose is called Gram's staining, named after

the 19th century Danish bacteriologist Hans Christian Gram. A Gram-positive bacterium has a single membrane inside a thick rigid structure cell wall, composed of a number of layers of peptidoglycan, to which anionic polymers (e.g. teichoic acids) are covalently attached. During the Gram stain procedure Gram-positive bacteria resist decolorisation and retains the colour of the initial stain, a blue/violet colour. A Gram-negative bacterium has a inner (cytoplasmic) membrane, a thin layer of peptidoglycan and an outer membrane. During the Gram stain procedure Gram-negative bacteria are decolorised with alcohol and counter stained with fuchin to give a pink colour [Hale et al., 1995, Sonenshein et al., 2002].

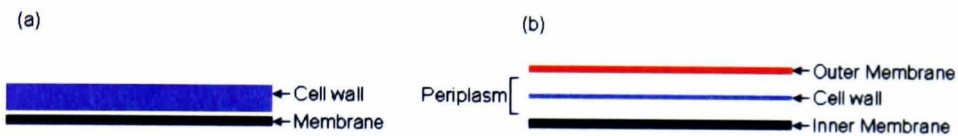


Figure 2.2: The cell envelope structures of (a) Gram-positive and (b) Gram-negative bacteria.

2.2.3 *Bacillus* species

Bacillus is a genus of Gram-positive, rod-shaped, aerobic or facultative, endospore-forming bacteria that can be found in almost any environment, but particularly in the soil. *Bacillus* species are low % GC content bacteria i.e. these bacteria have a low percentage of GC base pairs in their genome. Those that have a low GC content have more AT base pairs, which are not as tightly coupled as GC base pairs, and are therefore less stable. Survival is achieved in some low % GC bacteria through the formation of endospores, often triggered by a depletion of nutrients in their environment. An *endospore* is a dormant, tough, non-reproductive structure produced by a small number of bacteria from the Firmicute family. These spores can survive for many years in the dormant state, but can be rapidly activated through a process called *germination*. The ability to produce endospores allows *Bacillus* species to survive under environmental stresses that would otherwise kill the bacteria, in-

cluding high temperature, radiation, desiccation, and toxic damage or destruction [Sonenshein et al., 2002] [Kenneth Todar University, 2005].

Bacillus species are important not only for their industrial uses in the production of enzymes, pharmaceuticals, detergents and other chemicals, but also because of the diversity of characteristics shown by the members of the genus *Bacillus*.

- ***Bacillus subtilis***

B. subtilis is found in water, soil, air and decomposing plant residue. Strain 168 of this non-pathogenic organism has been the focus of intense studies, and consequently is the best characterised Gram-positive bacterium. In combination with the model Gram-negative bacterium *Escherichia coli*, the evolutionary divergence of eubacteria into the Gram-positive and Gram-negative groups can be studied. In addition, *B. subtilis* and its close relatives are able to secrete enzymes into the culture medium at high concentrations. As a result, *B. subtilis* has been used for the study of protein secretion. In a commercial context, *B. subtilis* provides an essential source of industrial enzymes used in products such as household detergents, peptide antibiotics, as well as antifungals that are useful in agriculture for crop protection [Kunst et al., 1997] [Bornemann, <http://www.micron.ac.uk>, U.S. Environmental Protection Agency, <http://www.epa.gov>].

- ***Bacillus licheniformis***

B. licheniformis is a close relative of *B. subtilis*. It is commonly found in soil and plant material. In humans, it is known to cause food poisoning [de Boer et al., 1994]. Like *B. subtilis*, *B. licheniformis* secretes large quantities of proteins and consequently has many uses in industry for the large scale production of enzymes [Veith et al., 2004]. Unlike most other bacilli, *B. licheniformis* is a facultative anaerobe and may therefore be possible for this bacterium to grow in other ecological niches. These species are usually saprophytic, as they are able to breakdown complex polysaccharides, and therefore contributing substantially to nutrient cycling. Certain strains are capable of denitrification,

which may have a relatively small impact on environmental denitrification as bacilli generally persist in the soil as endospores [Rey et al., 2004] [US EPA, <http://www.epa.gov>].

- ***Bacillus halodurans***

It has been indicated that alkaliphilic *B. halodurans* is more closely related to *B. subtilis* than any other *Bacillus* strain. Like *B. subtilis* and *B. licheniformis*, *B. halodurans* has applications for the industrial production of enzymes [Takami et al., 2001, 2000] [Jamstec, <http://www.jamstec.go.jp>].

- ***Bacillus clausii***

In addition to having applications in the production of enzymes for the detergent industry, alkaliphilic *B. clausii* is an important human probiotic [Cenci et al., 2006]. Probiotics are live microbial food supplements use to maintain or improve intestinal microbial balance. Probiotic bacterial cultures stimulate the growth of favoured microbes, excluding potential pathogens, and so reinforces the natural defence mechanisms of the body [Saarela et al., 2000]. Some *B. clausii* strains have been found to release antimicrobial substances that have been shown to be active against other Gram-positive bacteria, particularly *Staphylococcus aureus* [Cenci et al., 2006].

- ***Bacillus anthracis***

The human pathogen *B. anthracis*, the causative agent of anthrax, is found in the soil of many countries throughout the world, including Asia, Africa, parts of Europe, as well as North and South America. Anthrax is mainly a disease of herbivorous mammals, but may also be contracted by other animals and some birds. Generally humans contract the disease through contact with infected animals or contaminated animal products, generally leading to cutaneous (skin) anthrax, or rarely ingestinal anthrax. Nowadays this organism is more widely associated with the threat posed by bioterrorism, due to the fact *B. anthracis*

can cause lethal inhalation anthrax [Andersen et al., 1996, Han et al., 2006, Read et al., 2003] [Health Protection Agency, <http://www.hpa.org.uk>].

The two plasmids, pXO1 and pXO2 encode the major virulence factors of *B. anthracis*. The plasmid pXO1 encodes components of the secreted exotoxins: protective antigen, lethal factor, and edema factor. The edema factor (a calmodulin-dependent adenylate cyclase) and the protective antigen (whose function is to allow the toxin to enter the host cell) combine to form the *edema toxin*, which inhibits neutrophil function. Neutrophils are a type of white blood cell important in immune response. *Lethal toxin* is made up of lethal factor (a zinc metalloprotease which stimulates the release of tumor necrosis factor α and interleukin-1 β from the macrophages) and protective antigen. In systemic anthrax, these factors are both partially responsible for sudden death. In essence, it is believed the exotoxins inhibit the immune response triggered by infection. The plasmid pXO2 is involved in the synthesis of the immune evading polyglutamyl capsule that inhibits the process of phagocytosis, an important defence mechanism whereby the bacteria is engulfed by the host's phagocytes (e.g. white blood cells). Three genes on the plasmid, *capA*, *capB* and *capC*, are responsible for the synthesis of the capsule. The most pronounced virulent effect is only expressed when both plasmids are present. For instance, the mildly virulent Pasteur strain lacks pXO1, hence it is able to synthesise the capsule but does not produce the anthrax toxins. The less virulent Sterne strain contains pXO1 only, so is able to produce the toxins but does not have a capsule [Dixon et al., 1999]. *B. anthracis* shows close phylogenetic relations to *B. cereus* and *B. thuringiensis* [Rasko et al., 2004].

- ***Bacillus cereus***

As with many organisms, a number of strains have been identified, but only three have had their genome completely sequenced: *B. cereus* (ATCC 14579), *B. cereus* (ATCC 10987) and *B. cereus* (E33L). These strains are considered soil-based opportunistic pathogens that mainly cause food poisoning. Differ-

ent types of toxins result in two types of gastrointestinal illness: diarrhoeal poisoning and emetic poisoning.

Diarrhoeal poisoning is presently known to be caused by three different enterotoxins:

- two protein complexes:
 - * haemolysin BL (HBL) - consists of three components, two lytic components (encoded by *hblC* and *hblD*) and a binding protein B (encoded by *hblA*).
 - * non-haemolytic enterotoxin (NHE) - consists of three components, encoded by three genes *nheA*, *nheB* and *nheC*.
- a single protein cytotoxin (CytK).

Emetic poisoning is caused by a toxin called emetic toxin or *cereulide*, a heat stable cyclic dodecadeptide encoded in *ces* genes. It has been shown that the genes responsible for this toxin are encoded on a plasmid similar to the pXO1 plasmid of *B. anthracis*.

B. cereus has also been implicated in local or systemic diseases including endophthalmitis, endocarditis, meningitis, osteomyelitis, and periodontitis. However, research is ongoing in unravelling the complex virulent pathways involved [Ehling-Schulz et al., 2006, Ivanova et al., 2003, Rasko et al., 2007]. Of these *B. cereus* pathogens, *B. cereus* (E33L) shows distinct similarity with the other pathogenic organisms *B. anthracis* and *B. thuringiensis* [Han et al., 2006].

Not all strains of *B. cereus* are pathogenic or soil-based. Other strains have been isolated from humans and dairy sources [Rasko et al., 2007]. Genome sequencing is not complete for these strains [NCBI, <http://www.ncbi.nih.gov>].

• *Bacillus thuringiensis*

B. thuringiensis is an insect pathogen that has been widely used as a biological insecticide. Products based on this bacteria are produced on a commercial scale

for use in the agricultural industry for pest control. The toxicity of this bacteria is due to the production of a δ -endotoxin. An endotoxin is a natural toxin produced in the mature cell during sporulation. This toxin is released when the mature cell lyses to release the endospores. Different strains of *B. thuringiensis* produce δ -endotoxins with different host specificities and consequently can be targeted to different insect pests. The *cry* genes responsible for the toxicity of *B. thuringiensis* have also been introduced into plants through genetic engineering, consequently leading to plant insect resistance [Hale et al., 1995, Rasko et al., 2004].

2.2.4 Mechanisms of secretion

The study of protein secretion not only includes the secreted proteins themselves, but also those proteins which form the secretory machinery (a.k.a. the translocase) used to export the proteins across the cytoplasmic membrane and cell wall. For a protein to be transported from its site of synthesis, the cytoplasm, into the culture medium, it first needs to be transported to the cell membrane. Secretory proteins generally incorporate a short amino-terminal sequence called a *signal peptide*, which acts as a targeting signal to secretory machinery. Once they are targeted to the translocase, the secretory proteins are either exported into the growth medium or retained at an extracytoplasmic location (i.e. cell membrane or cell wall) [Tjalsma et al., 2004, 2000, van Wely et al., 2001].

Within *B. subtilis*, at least four different protein transport pathways are known to exist. These allow proteins to be steered to at least five different subcellular locations depending on the presence or absence of an amino-terminal signal peptide and specific retention signals [Boekhorst et al., 2005, Tjalsma et al., 2004, 2000]. These protein transport pathways (highlighted in figure 2.3), as well as the mechanisms of protein retention, are discussed.

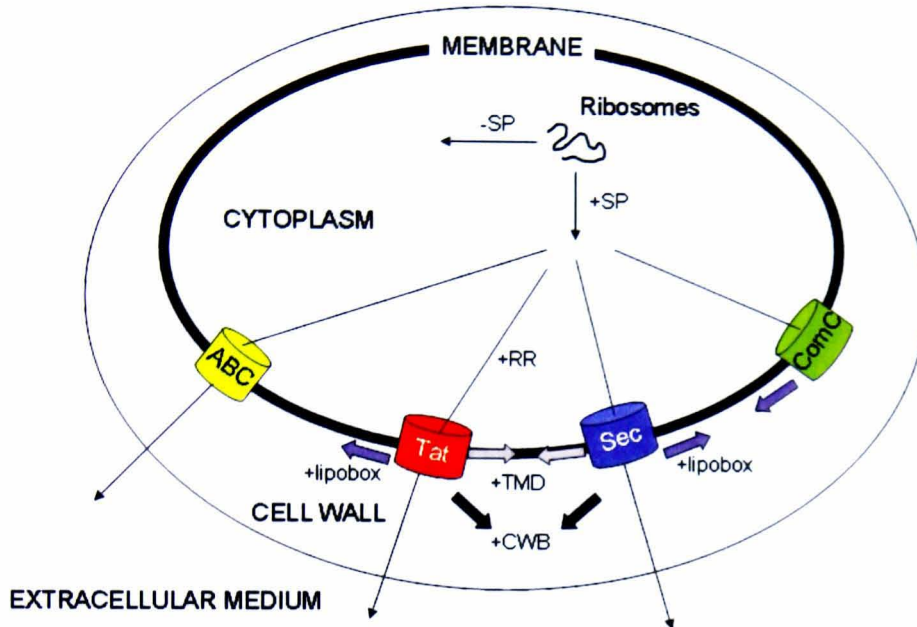


Figure 2.3: Proteins synthesised by the ribosomes with a signal peptide (+SP) are transported to different extracytoplasmic locations surrounding the cell. Proteins with transmembrane domains (+TMD) are inserted into the membrane via the general secretory (Sec) pathway or possibly the twin-arginine (+RR) translocation (Tat) pathway. Lipid-modified proteins (+lipobox) may attach themselves to the outer membrane surface via the Sec or Tat pathway; prepilins exported by the Com pathway may also function in this way. Proteins exported via the Sec or Tat pathway may also be retained in the cell wall if the protein contains cell wall-binding repeats (+CWB). Proteins destined for the extracellular medium can be secreted via the Sec or Tat pathway, or by ABC transporters. Proteins without a signal peptide (-SP) remain in the cytoplasm or associated with the membrane. Adapted from Tjalsma et al. [2000].

2.2.4.1 Amino-terminal signal peptides

Amino-terminal signal peptides (sometimes referred to as prepeptides) are essentially export signals that are recognised by cytoplasmic chaperones. These chaperones guide the proteins to the translocation machinery for secretion [Campo et al., 2004, van Wely et al., 2001]. The structure of the amino-terminal signal peptides can be split into three distinct domains:

- The amino-terminal *N-domain* contains at least one positively charged arginine or lysine residue; although this feature does not appear to be needed for protein export. It has been suggested that the positively charged N-domain plays a role in translocation, interacting with the translocation machinery and negative phospholipids in the lipid bilayer of the membrane during this process.
- The *H-domain* consists of hydrophobic residues that appear to form an alpha-helical conformation in the membrane. In the middle of this domain, glycine or proline residues are often found. These helix-breaking residues may allow the signal peptide to adopt a hairpin-like structure that inserts into the membrane. Additional helix-breaking residues at the end of the H-domain possibly aid the cleavage of the signal peptide from the protein by a specific signal peptidase (SPase) during or shortly after translocation.
- The cleavage site for the SPase is found in the *C-domain*. Cleavage at this site results in the protein being released from the membrane and folding into its native conformation. The signal peptide is then degraded by signal peptide peptidases (SPPases) and removed from the membrane.

Despite the similarity in amino-terminal signal peptides, small differences in their structure affects the pathway used in protein export and the final destination of the protein. These differences result in cleavage by one of a number of different SPases. In *B. subtilis*, four different types of signal peptides have been identified based on SPase recognition [Tjalsma et al., 2004, 2000, van Wely et al., 2001].

- **Secretory (Sec-type) signal peptides**

Sec-type signal peptides are regarded as the major class of signal peptides, cleaved from the mature protein (during or after translocation) by a type I SPase. The resulting processed proteins can potentially be secreted into the extracellular medium or alternatively retained at the membrane or cell wall (figure 2.4a). A subtype of this group contains a twin arginine motif (RR motif), which directs proteins to an alternative pathway, termed the Tat pathway [Tjalsma et al., 2000, van Wely et al., 2001] (figure 2.4b).



Figure 2.4: The structure of (a) Sec-type signal peptides and (b) twin-arginine signal peptides. The numbers indicate the average length of the signal peptide in terms of the number of amino acid residues. Adapted from Tjalsma et al. [2000].

- **Lipoprotein signal peptides**

The presence of a well-conserved lipobox in prelipoproteins distinguishes lipoproteins from Sec-type proteins (figure 2.5). The lipobox acts as a recognition signal for lipid-modification. The lipobox contains an essential cysteine residue. It is through the lipid-modification of this cysteine residue that the lipoprotein remains attached to the membrane. The signal peptides of prelipoproteins are cleaved by type II SPase [Tjalsma et al., 2004, 2000, van Wely et al., 2001].

As with secretory proteins, an RR motif has been identified in some prelipoproteins, hence some lipoproteins are predicted to be transported via the Tat pathway as opposed to the Sec pathway [McDonough et al., 2005]. In view of the

small number of lipoproteins translocated via the Tat pathway, the significance of this pathway with respect to lipoproteins is unclear [Rezwan et al., 2007].

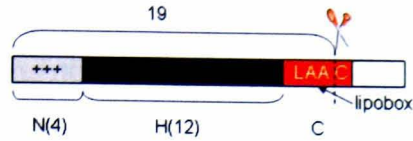


Figure 2.5: The structure of lipoprotein signal peptides. The numbers indicate the average length of the signal peptide in terms of the number of amino acid residues. Adapted from Tjalsma et al. [2000].

- **Bacteriocins and pheromones signal peptides**

The signal peptides of bacteriocins and pheromones are composed of N- and C-domains only (figure 2.6); they lack a typical hydrophobic domain. Therefore regular predictive algorithms cannot be applied. These proteins are exported by ABC transporters. Removal of the signal peptide from the mature protein is carried out either by a subunit of the ABC transporter that is specific to a particular bacteriocin or pheromone, or by specific SPases [Biemans-Oldehinkel et al., 2006, Tjalsma et al., 2004, 2000].

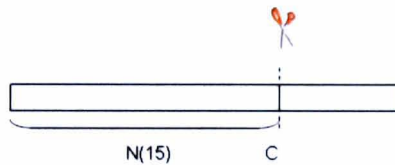


Figure 2.6: The structure of signal peptides of bacteriocins and pheromones. The number indicates the average length of the signal peptide in terms of the number of amino acid residues. Adapted from Tjalsma et al. [2000].

- **Prepilin-like signal peptides**

The C-domain of prepilin-like signal peptides is located between the N- and H-domains (figure 2.7). Therefore after cleavage by ComC (at the cytoplasmic side of the membrane), the hydrophobic H-domain remains attached to the

mature protein, allowing transport through the membrane [Chen and Dubnau, 2004, Tjalsma et al., 2004, 2000].

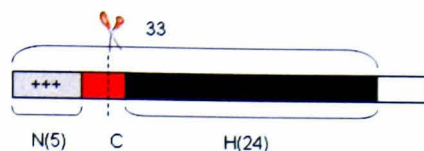


Figure 2.7: The structure of prepilin-like signal peptides. The numbers indicate the average length of the signal peptide in terms of the number of amino acid residues. Adapted from Tjalsma et al. [2000].

2.2.4.2 Translocation machinery

As illustrated in figure 2.3, proteins can be targeted to different translocation pathways depending on the structure of their amino-terminal signal peptide, i.e. the general secretory (Sec) pathway, the twin-arginine translocation (Tat) pathway, a prepilin-like pathway and via ABC transporters [Tjalsma et al., 2000]. Interestingly, a potentially novel pathway may exist that has previously been uncharacterised. ESAT-6 is a secreted protein of unknown function, important in the virulence of *Mycobacteria tuberculosis*. However it lacks a signal peptide, and therefore it must utilise an alternative secretion pathway. Advances have been made in recent years to understand the mechanisms of the pathway involved. Homologues of ESAT-6 have also been identified computationally in other bacteria, including *B. subtilis* and *B. anthracis* [Brodin et al., 2004, Pallen, 2002].

These membrane transport systems generally consist of one or more protein components, where at least one of these proteins contains multiple transmembrane alpha-helical segments allowing it to span the membrane. The translocation of proteins through the membrane channel is often driven by ATP hydrolysis [Sonenshein et al., 2002].

- **General secretory (Sec) pathway**

The main pathway through which proteins are transported from bacteria is the Sec pathway. Extensive research has been conducted into the functionality of this highly conserved pathway in *E. coli* and *B. subtilis* [Tjalsma et al., 2000, van Wely et al., 2001]. Extracellular, membrane and cell wall proteins are translocated via the Sec pathway through a channel formed by the membrane-embedded protein complex SecYEG in an unfolded or partially folded conformation. The translocation motor SecA, a peripherally bound ATPase, drives the translocation process. SecA interacts with the preprotein and subsequently undergoes repeated binding and hydrolysis of ATP, combined with conformational changes, eventually resulting in translocation [Campo et al., 2004]. The signal peptide can be cleaved by SPase I, or SPase II in the case of lipoproteins, before complete translocation of the mature protein [Berks et al., 2005].

- **Twin-arginine translocation (Tat) pathway**

The Tat pathway is important in a number of bacterial processes, including energy metabolism, cell wall biosynthesis and pathogenesis. This pathway is responsible for the export of folded preproteins. It consists of three integral membrane proteins TatA, TatB and TatC. Studies on *E. coli* indicate that TatA forms the channel through which proteins pass, and TatBC is responsible for signal peptide recognition. The mechanism begins with the binding of the preprotein to the TatBC complex, specifically binding to TatC. The TatBC-preprotein complex then associates with TatA for transport across the membrane. It is thought that the signal peptide is then cleaved by SPase I, or SPase II in the case of lipoproteins, only after the complete translocation of the mature protein [Berks et al., 2005, Tjalsma et al., 2004].

- **ATP-binding cassette (ABC) transporters**

ATP binding cassette (ABC) transporters are integral membrane proteins that transport proteins across biological membranes. ABC transporters, like other

secretion pathways, are important for uptake of nutrients and elimination of waste products, energy generation, and cell signalling. In order to secrete proteins, ABC transporters require a minimum of four domains: two transmembrane domains that form the protein binding sites and provide specificity, and two ATP binding domains that provide a mechanism for driving the translocation of the protein via ATP catalysis [Biemans-Oldehinkel et al., 2006, Linton, 2007].

- **Type IV prepilin-like pathway**

Type IV prepilin-like proteins are involved in the development of genetic competence. *Competence* is a state in which a bacteria is able to take up DNA from another bacterium; this process is called *transformation*. These prepilin-like proteins are exported via an alternative pathway to those previously discussed, called the ComC pathway. ComC is an integral membrane protein that is therefore bifunctional. In terms of secretion of prepilin-like proteins, it is required for the assembly and anchoring of the pilin-like structures to the membrane, requiring cleavage of the signal peptide. The phenylalanine residue at position 1 in the resulting mature protein is then methylated by ComC [Chen and Dubnau, 2004, Tjalsma et al., 2000].

2.2.4.3 Retention mechanisms

Despite the presence of an amino-terminal signal peptide-like sequences, not all of these proteins are exported into the environment. Many proteins are retained at the cell surface through additional retention mechanisms.

- **Transmembrane domains**

Some proteins contain additional hydrophobic regions that can integrate into the membrane and span across the membrane. These membrane proteins are only partially translocated; they embed themselves in the membrane once they have been released from the translocation process. Transmembrane proteins are

found in transporter complexes, channels and enzymes, all of which have a role in protein secretion [Sonenshein et al., 2002, Tjalsma et al., 2004].

- Lipid-modifications

Lipoproteins remain anchored to the membrane through the lipid-modified cysteine residue, most likely through hydrophobic interaction. Lipoproteins are associated with many different functions important in cell survival, including adhesion and invasion, cell wall synthesis, nutrient uptake, degradative processes, and sensing and transmembrane signalling [Sutcliffe and Russell, 1995].

The mechanism through which prelipoproteins undergo lipid-modification is the responsibility of two consecutive enzymatic reactions involving diacylglycerol transferase and SPase II. These enzymes are membrane-integral proteins (figure 2.8). Cleavage of the signal peptide from the mature protein is initiated by diacylglycerol transferase, which targets the cysteine residue in the preproteins, adding a diacylglycerol group to the thiol group of cysteine. The signal peptide is subsequently cleaved by SPase II.

Research suggests that lipoproteins play an important role in pathogenicity. Inactivation of the genes involved in lipoprotein biosynthesis reduces the virulence of Gram-positive pathogens [Rezwan et al., 2007].

- Cell wall binding

In Gram-positive bacteria, the cell wall functions as a protective layer, guarding the cell against osmotic damage, as well as defining the cell's shape, but also providing mechanisms through which the bacteria can interact with its environment. Pathogenic bacteria can alter their cell wall structure and function to adapt to changes in their environment.

The main component of the cell wall is peptidoglycan, which contains additional molecules that are covalently and non-covalently attached, including teichoic acids, polysaccharides and proteins. The result of all these interactions provides the cell wall with properties specific to the species and strain, greatly

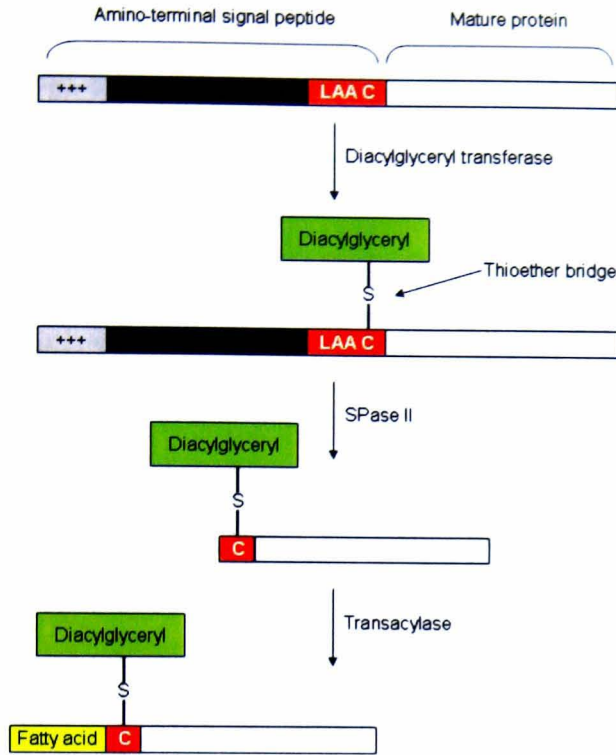


Figure 2.8: The steps involved in processing lipoproteins prior to the attachment to the outer surface of the membrane. The covalent attachment of a lipid residue to the conserved cysteine is the result of three consecutive enzymatic reactions involving: diacylglyceryl transferase, SPase II and transacylase. Adapted from Rezwan et al. [2007].

contributing to the virulence, host interactions, and disease symptom development and outcome of pathogens [Marraffini et al., 2006].

Just as a protein (with the necessary domains) can bind to the membrane, a protein may also possess the potential to be retained at the cell wall, via covalent or non-covalent mechanisms. Of particular interest are those cell wall proteins that are covalently anchored to the surface. These surface proteins are distinguished by the presence of a cell wall sorting signal at the C-terminal end of a protein. The cell wall sorting signal is composed of a short pentapeptide motif (LPXTG) followed by hydrophobic region and a positively charged tail.

It is approximately 30 to 40 residues long (figure 2.9) [Cossart and Jonquières, 2000, Marraffini et al., 2006].

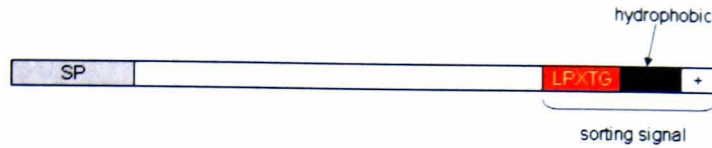


Figure 2.9: The structure of surface protein precursors, containing a signal peptide (SP) at the N-terminal and the sorting signal at C-terminal. The sorting signal consists of an LPXTG motif, a hydrophobic domain and a positively charged tail.

The mechanism of covalent attachment to the cell wall is catalysed by transpeptidation enzymes known as *sortases*. The steps involved are highlighted in figure 2.10. Following the translocation of a surface protein precursor across the membrane and the cleavage of the signal peptide by SPases, the LPXTG cell wall sorting signal retains the mature protein within the secretory pathway. The sortase enzymes then catalyse a transpeptidation reaction; the LPXTG motif is targeted by the sortase, promoting cleavage of the sorting signal from the protein and the subsequent anchoring of the protein to the cell wall by attacking a bridge of peptidoglycan precursors [Navarre and Schneewind, 1999].

2.3 Data integration, e-science and Grid technology

With the advent of the omic revolution there is a real need for the development of analytical tools that can process the vast quantity of data being generated, and ensure the sharing and integration of this data to enable better-informed or faster decisions, and minimise the rediscovery of already known facts [Hey and Trefethen, 2005]. However, within bioinformatics, integration is a well-known problem [Hull et al., 2006, Stevens et al., 2007]. Data analysis is often based on data from different sources, which may include gene and protein sequence data, measurement data from

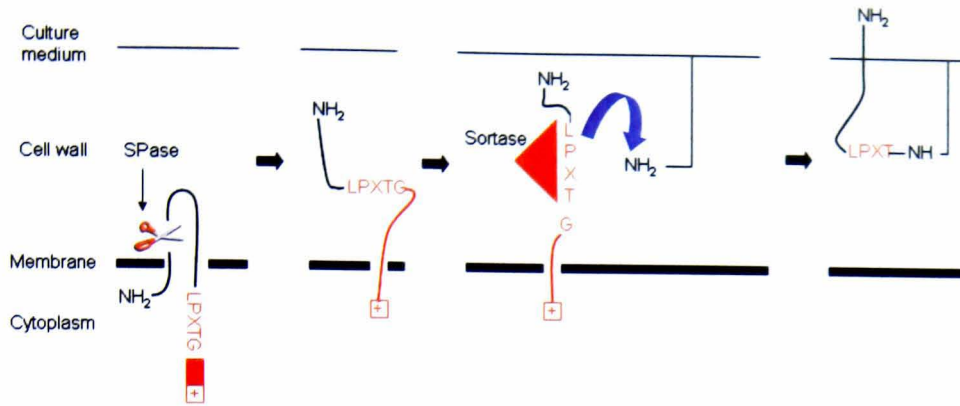


Figure 2.10: The mechanism of covalent attachment of a protein to the surface of the cell wall through an LPXTG motif. Adapted from Cossart and Jonquière [2000].

different types of instruments, and so forth. Along with the data itself, metadata can be extremely informative. *Metadata* is data about data that describes, explains and locates a data resource, facilitating in the retrieval, use, and management of the resource [NISO, www.niso.org]. Of particular interest to biologists is a subtype of metadata known as *provenance*. Provenance provides information regarding the process of biological experiments, the purpose and results of experiments, along with annotations and notes provided by the scientist about the experiment. This metadata is essential for others to be able to validate and verify the integrity of the experiment [Greenwood et al., 2003].

Data and metadata need to be stored and accessed efficiently from remote locations by different groups. Furthermore, tools are needed to compare, classify and analyse this data to extract meaning and to generate further hypotheses that can be tested in the lab [Gray et al., 2005, Srivastava and Velegrakis, 2007]. This process is difficult since tools and data sources are distributed all over the world, in different formats, and of variable quality. Currently many life scientists work round the problem by simply cutting and pasting between many different Web pages. However this temporary solution does not scale and is not easy to apply to large, omic-scale datasets when data about thousands of proteins needs to be handled.

There are therefore still many barriers to be overcome to improve the sharing and integration of data. Dealing with the three problems - heterogeneity, distribution and autonomy - often forms a significant part of the work load of a bioinformatician [Hernandez and Kambhampati, 2004, Karasavvas et al., 2004, Stevens et al., 2007].

1. *Heterogeneity*

Bioinformatics tools and datasets tend to be highly heterogeneous in content and structure. When biological and clinical data is represented electronically, a multitude of formats are used presenting a barrier to open information exchange within the scientific community. Standardised knowledge representations can help to address this problem, but so far uniform standardisation has been impossible to achieve on a large scale.

2. *Distribution*

Datasets are often widely geographically dispersed, being distributed on multiple servers in multiple locations. However, using distributed resources has associated problems: networks can fail, data transfer is slow and expensive, and a crucial requirement is security.

3. *Autonomy*

Tools and datasets are often maintained and deployed by individual scientists within their own laboratories. As a result, reliability and robustness of data are often a concern [Hernandez and Kambhampati, 2004, Karasavvas et al., 2004, Stevens et al., 2007].

2.3.1 e-Science and the Grid

e-Science refers to a science in which scientists from around the world collaboratively share data and distributed computing resources [Hey and Trefethen, 2003, 2005]. In order to provide a platform for e-science, technologies such as the Web, Web services and the Grid are used. The Web is a service for sharing information over the Internet [W3C, <http://www.w3.org>] and the Grid is a service for sharing computer power and

data storage capacity over the Internet [Foster et al., 2001]. Web services are services for sharing applications/algorithms over the Internet [W3C, <http://www.w3.org>] that can be combined into more complex systems through building workflows.

The ultimate aim of the Grid is to create a large computational resource utilising the global network of computers, belonging to different people and institutions, including desktop PCs and workstations, mainframes and supercomputers, but also incorporating data vaults and instruments such as meteorological sensors and visualisation devices. The advantage of this approach is that it removes the need to install programs locally or download data, allowing remote execution of a program or analytical tool, using a remote machine where the resource is located, or if this machine is busy, the Grid is designed to copy the resource to another computer or cluster of computers which are idle for execution. Furthermore, the Grid would allow interactive collaborative analysis, by networking computers to give the feel of a local network [Clery, 2006][CERN, <http://gridcafe.web.cern.ch>].

The Grid is based upon coordinated resource sharing and problem solving in dynamic, multi-organisational virtual organisations. A virtual organisation consists of a set of individuals and/or institutions defined by sharing rules, in which resource providers and consumers state clearly what is shared and the conditions under which sharing occurs. Essentially, the Grid is the technology to provide transparent data sharing between organisational and geographic borders. The Grid tackles many concerns and requirements not addressed by current distributed technologies, such as the need for flexibility and control over how a variety of resources (programs, files and data or computers, sensors and networks) are shared and used, as well as issues concerning quality of service, scheduling and accounting [Foster et al., 2001][CERN, <http://gridcafe.web.cern.ch>].

In reality there is no single Grid. Many Grids exist including Grids that are either private or public, regional, national or even global, all-purpose or specific to a scientific problem [CERN, <http://gridcafe.web.cern.ch>].

To enable the construction of Grids, complex systems of software and services are required [CERN, <http://gridcafe.web.cern.ch>]. Over recent years the Grid community

has succeeded in developing protocols, services and tools that provide an answer to many of the challenges being faced in this area [Foster et al., 2001].

One of the areas where the interests of scientists and businessmen converge is *standards*. Common open standards are required to ensure that development of the Grid is constructive, and that applications that can run on one Grid can run on another. This is an important issue with the Grid, as the Grid's main purpose is the sharing of resources. Like the Internet Engineering Task Force (IETF) for the Internet and the World Wide Web Consortium (W3C) for the Web, the Open Grid Forum¹ is responsible for developing standards specific to the Grid. Through this forum, a standard known as Open Grid Services Architecture (OGSA) is being developed that already provides a key reference for future Grid development projects. OGSA seeks to harmonise the development of the Globus Toolkit, a software package for creating grids. The toolkit provides a set of software tools to implement the basic services and capabilities required to construct a computational Grid, such as security, resource location, resource management, and communications. The Globus toolkit makes use of Web services, an additional standard for services offered over the World Wide Web [Clery, 2006, Foster, 2003][CERN, <http://gridcafe.web.cern.ch>].

2.3.1.1 Architecture

The architecture of the Grid is often described in terms of layers, each having a specific function (figure 2.11.a) [Foster, 2003][CERN, <http://gridcafe.web.cern.ch>]; the lower layers are focused on computers and networks, whereas the higher layers are focused on the user:

- The *network layer* connects the resources in the Grid.
- The *resource layer* is composed of the actual Grid resources (e.g. computers, sensors) that can be directly connected to the network.
- The *middleware layer* provides the tools that enable the various Grid resources to function on the Grid; this is the 'brain' of the Grid.

¹Open Grid Forum Website: <http://www.ogf.org/>

- The *application and serviceware layer* is the layer that users interact with, incorporating the user applications, and often *serviceware* that is responsible for measuring a user's Grid usage, tracking resource providers and users of those resources.

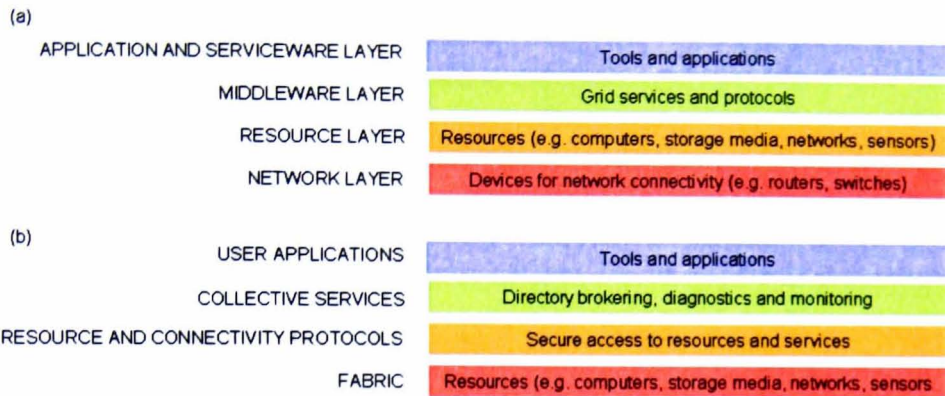


Figure 2.11: The Grid architecture in terms of (a) tools and resources (adapted from CERN, <http://gridcafe.web.cern.ch>) and (b) the protocols and application programming interfaces (adapted from Foster [2003]).

An application that was originally implemented to run on a standalone computer needs to be modified to run across the Grid, in order to invoke all the correct services and protocols, just as it would otherwise have to be adapted to run on the Web. A redefined model that is shown in figure 2.11.b, illustrating the Grid architecture in the context of protocols:

- The *fabric* defines the physical infrastructure of the Grid, including computers and the communication network.
- *Resource and connectivity protocols* are responsible for the network transactions between the different resources on the Grid (commonly obtained through using the Globus Toolkit):
 - *Communication protocols* to allow resources to exchange data over the Internet.

- *Authentication protocols* to verify the identity of users and the resources.
- *Collective services* are responsible for tracking available resources, brokering resources (i.e. mediating between resource providers and users), monitoring and diagnosing problems, ensuring multiple copies of data are available at different locations, and tracking who is allowed to do what, when. In order to provide this functionality additional protocols are used:
 - *Information protocols* obtain information about the structure and state of the resources on the Grid.
 - *Management protocols* negotiate access to resources in a uniform way.
- *User applications* once again represent the users view on the Grid.

In this latter representation (figure 2.11.b), the middleware layer has been split based on resource and connectivity protocols, and the collective services, whereas the network layer and resource layer have been fused into the new fabric layer.

Each layer in the architecture is needed in order to be able to execute an application over the Grid. For instance, a Grid application whose function is to analyse the data stored in several files follows several steps:

- Obtain authentication in order to open the files (using resource and connectivity protocols).
 - Determine the current locations on the Grid of the appropriate files, as well as the most convenient location of the computational resources (uses collective services).
 - Extract the data, analyse, and return the results (uses resource and connectivity protocols).
 - Monitor the progress of the data transfers and analysis, notifying the user when the analysis is complete, and responding to any failure (uses collective services)
- [Foster, 2003][CERN, <http://gridcafe.web.cern.ch>].

2.3.2 Services on the Web

2.3.2.1 Soap-based Web services

Web services are the next generation of Web-based technology for exchanging information. Web services allow Web-based applications to be built using any platform and programming language. Due to their modularity and flexibility, Web services are ideal for integrating applications, particularly in making heterogeneous resources interoperable. Following the deployment of a Web service, other applications and Web services can discover and invoke that service [Booth et al., 2004, Gottschalk et al., 2002, Wang et al., 2004].

Exposing bioinformatics applications as Web services eliminate the problems associated with using several different applications that need to be installed locally. The execution of these applications can be coordinated by integrating Web services into workflows [Altunay et al., 2004]. A number of tools are now available allowing biologists to easily compose and enact workflows, without the need for technical expertise in computing. An example tool is Taverna, which provides a workflow markup language (SCUFL) along with software to construct workflows using distributed technology [Oinn et al., 2004].

SOAP-based Web services are a common type of Web service. The components involved in finding, publishing and binding to a Web service can be described by considering the service-oriented architecture of a Web service, as shown in figure 2.12. This architecture consists of a service provider, service requester and service registry:

- A *service provider* creates a Web service, and the corresponding service description may be published within a service registry (or obtained by simply emailing the service requester).
- The *service registry* then provides the service requester with the service description and a Uniform Resource Locator (URL) indicating the location of the service.

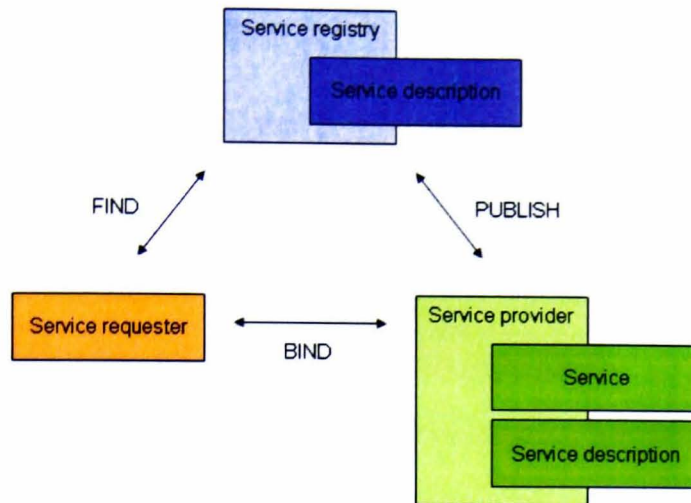


Figure 2.12: SOAP-based Web services architecture. Adapted from Gottschalk et al. [2002].

- The *service requester* can then use this information to bind to the service and execute it [Booth et al., 2004, Gottschalk et al., 2002].

Web services utilise a number of standardised protocols and application programming interfaces (APIs) enabling users to locate and use the operations provided. This is highlighted in the Web services programming stack (figure 2.13). The programming stack is broken down into a number of levels:

- The *network* is the bottom level over which a Web service is made available, and is often based on the HyperText Transport Protocol (HTTP) protocol.
- *XML-based messaging* is based on SOAP that defines a mechanism enabling communication between Web services, providing a way of exchanging information between service requesters, service registries and service providers.
- A *service description* describing the operation and location of the available Web services is provided through the Web Service Description Language (WSDL), a formal, computer-readable XML-based language to describe Web services. A

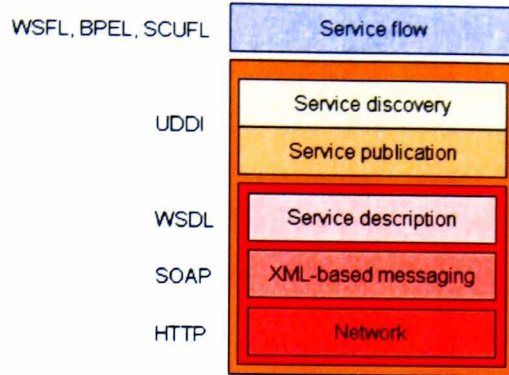


Figure 2.13: SOAP-based Web services programming stack. Adapted from Gottschalk et al. [2002].

services WSDL file is all that is required to create an application that communicates with a particular Web service.

- *Service publication* and hence *service discovery* can be done through a Universal Description, Discovery, and Integration (UDDI) registry that stores descriptions about services offered by an institution in a common XML format to allow clients to find a WSDL for a particular Web service and invoke the service.
- The *service flow* layer enables Web services to be integrated into workflows. BPEL4WS [Andrews et al., 2003] and SCUFL [Oinn et al., 2004] are *potential* standards that could be used at this level [Gottschalk et al., 2002, Wang et al., 2004].

The network, XML-based messaging and service description layers are required in order to have interoperable Web services [Gottschalk et al., 2002]. Apache Axis² and Codehaus XFire³ are open source frameworks providing a foundation on which to build Java Web services. They support the important Web service standards, including SOAP over HTTP and WSDL. The remaining top-level layers are not necessary, but provide improved means of finding, binding and integrating Web services

²Apache Axis Website: <http://ws.apache.org/axis/>

³Codehaus XFire Website: <http://xfire.codehaus.org/>

[Gottschalk et al., 2002].

There are two ways in which SOAP messages are structured: Remote Procedure Call (RPC) and Document style. In RPC style, the contents of the body of the SOAP message must conform to a structure that indicates the method name and contains a set of parameters. In Document style, the contents of the body of the SOAP message is structured in any desired fashion [Curbera et al., 2002].

2.3.2.2 REST services

”Representational State Transfer is intended to evoke an image of how a well-designed Web application behaves: a network of web pages (a virtual state-machine), where the user progresses through an application by selecting links (state transitions), resulting in the next page (representing the next state of the application) being transferred to the user and rendered for their use” [Fielding, 2000].

Representational State Transfer (REST) is an architectural style used to describe network systems, such as the Web. Therefore, it is not a standard on its own, unlike SOAP which is a general protocol that can be used as part of different architectures [zur Muehlen et al., 2005].

It is based on the concept of a *resource*, where a resource is defined as anything that has a Universal Resource Identifier (URI). REST services interface with HTTP via the following commands:

- HTTP GET to retrieve a resource *representation* from a URI.
- HTTP DELETE to remove a resource representation.
- HTTP POST to update/create a resource representation.
- HTTP PUT to create a resource representation [Fielding and Taylor, 2002, zur Muehlen et al., 2005].

The differences between REST and SOAP is summarised nicely in Swenson [2005], which states:

"REST says that everything is a resource, and has a distinct Web address. In its purist form, it only supports HTTP, and the only supported operations are GET, PUT, POST and DELETE. Getting a resource gives you an XML structure. Web locations are objects, and you send the operation to the resource. SOAP, on the other hand, uses an address to fix the location of an API call. Each operation has an address, and if you want to talk about an object, you pass a reference (or a copy) of the object to the operation".

REST vs. SOAP is equivalent to object-oriented vs. procedure-oriented [Swenson, 2005]. Compared to SOAP-based approaches, REST is a lighter-weight and less feature-rich approach to Web services due to the limited number of operations and the unified address schema, as well as the potential scalability of REST-based systems. In contrast, SOAP has tight coupling of operations, allowing applications to be tested and debugged before deployment [zur Muehlen et al., 2005].

There is now a growing interest in REST services due to benefits it offers in terms of scalability, performance, security, reliability and extensibility [Fielding, 2000, Fielding and Taylor, 2002]. A compromise is now on offer with the incorporation of REST principles into the standards and guidelines of the new version of SOAP [Mitra and Lafon, 2007].

2.3.3 Taverna and *my*Grid

The *my*Grid project has developed tools that allow researchers to develop and execute their own workflows. Its aim is to provide free software to support the development of workflows. At the core of development is the Taverna workbench. Using Taverna, distributed programs and data sources can be accessed. Workflows can be graphically built using a huge number of services, edited, browsed and then executed using the built in workflow enactment engine, Freefluo [Hull et al., 2006][Taverna, <http://taverna.sourceforge.net>].

Freefluo coordinates the execution of parallel and sequential activities in the work-

flow, and supports data iteration and nested workflows. It is integrated into the Taverna workbench, but is also available as a desktop tool and Web service. Workflow definitions can therefore be executed by direct submission to the workflow enactor [Taverna, <http://taverna.sourceforge.net>]. Workflows are defined using the language called Simple Conceptual Unified Flow Language (SCUFL) [Oinn et al., 2004].

Within Taverna, processes can be added to a workflow in many ways, depending on the nature of the component. Local components, with existing Java APIs, can be imported into a Taverna workflow using the annotations derived by the API Consumer tool for the associated Java libraries [Taverna, <http://taverna.sourceforge.net>]. Remote components, on the other-hand, can be incorporated into the workflow using a multitude of methods.

- Assess to the operations of a standard SOAP Web service is provided through the *Web service client*.
- *Soaplab* is a tool that provides a way to automatically generate and deploy Web services to execute existing command-line analysis programs. Java implementation classes and deployment descriptors for the services are created using Apache Axis. The advantage of Soaplab is that it provides a unified access route to remote computational methods that may have limited or no user interface defined. (As a sub-project to Soaplab, Gowlab can be used to generate Web services on top of existing Web resources by providing data from a third-party Web page as a Web service) [Senger et al., 2003][Soaplab, <http://www.ebi.ac.uk>].
- *Styx Grid Service* is a method to wrap command-line programs to enable execution of these programs over the Internet. A Styx service is not a Web service, but like Web services, Styx Grid services can be composed into workflows using either simple shell scripts or graphical tools. The main advantage of Styx lies in its architecture; the Styx Grid service architecture provides a way of streaming data peer to peer (i.e. from service to service). Styx is therefore useful when using large datasets as it removes the need to transfer data through the client [Blower et al., 2006].

In addition, there are a number of local Java operations provided as default within Taverna that can be used for common generic tasks (e.g. to flatten lists). Also, during workflow construction it is often the case that the output of one processor does not have the correct format for the input of the next service. For simple cases writing a simple script to plug into the workflow as a non-service component may be appropriate. These Beanshell scripts are interpreted in Java [Taverna, <http://taverna.sourceforge.net>].

Furthermore, Taverna also provides ways of managing the resulting data, such as the ability to export to Microsoft Excel [Taverna, <http://taverna.sourceforge.net>]. It also records provenance, of which there are two categories: derivation data and annotations. Derivation data describes what initial data was used to obtain a result, and how the initial data was transformed into the result. In terms of an *in-silico* experiment, the derivation data is about which services were used and how they transformed the initial inputs into the overall result. Annotation data provides background information like who performed an experiment and when, were any comments supplied on the specific methods and materials used [Greenwood et al., 2003].

Provenance has a variety of uses. Firstly, another e-scientist should be able to repeat the experiment and verify the results with the information provided. Provenance data can also provide management information about what has been done in a virtual organisation: who last used a particular workflow, and in what context, what have others done with similar data.

2.3.4 Microbase

The ability to extract information about an organism's phenotype and metabolic potential by analysis of its genome sequence requires genome databases to be searched, and comparison and analysis to be performed. A number of tools have been developed to perform sequence comparison. Probably the most popular of these is BLAST [Altschul et al., 1990]. BLAST searches protein and nucleotide (DNA) databases to identify pairwise sequence similarities by local alignment between a query sequence

and each sequence in a database.

However, due to ever increasing volumes of genome data, genomic comparison and analysis has become a data-intensive and compute-intensive task. One possible solution is to employ Grid computing to build integrated genome databases and powerful computing environments for genomic data processing.

The Microbase project has developed a Grid-based system to provide a foundation for large-scale genome comparison and analysis by utilising the data and computing resources on the Grid [Sun et al., 2005] [Microbase, <http://www.microbase.org.uk/>]. The system stores pre-computed datasets of all-against-all microbial genome comparisons generated by a variety of genome comparison tools. It provides a scalable computing environment in which computationally intensive genome comparison and analysis can be performed. The system has been used to carry out a variety of applications including an all-against-all search for gene homologues (including orthologues and paralogues).

The EMBL database is dynamically integrated with the pre-computed dataset; through a Web service based notification system, newly published genomes in EMBL can be automatically added and compared against all existing genomes, thereby updating the pre-computed dataset. An additional Web service interface also provides remote users with access to this resource [Sun et al., 2005].

2.4 Biological network analysis

One aim of systems biology is to understand biology in the global scale of protein interaction networks. However, the interaction between proteins is not well defined and is used in various contexts when describing interaction networks, to mean proteins that physically interact, through to proteins that may be involved in some kind of common functional or metabolic process or transient protein complex, but may not necessarily physically interact. These global networks of proteins and their 'interactions' are called interactomes and can be defined using a variety of experimental techniques.

Interactomes not only provide a framework to analyse and experimentally verify theories concerning a pathway, but they can also reveal unknown properties and unanticipated consequences of different pathway configurations. Furthermore, they also provide a means of managing the complexity of a vast number of cellular components and interactions. One of the major applications of interactomes is in the prediction of protein function by the so called 'guilt-by-association' procedure [Cusick et al., 2005]. Guilt-by-association refers to the process of inferring the function of an uncharacterised protein on the basis it interacts with one of known function [Oliver, 2000].

2.4.1 Interactome modelling

A number of studies have been undertaken for analysing protein-protein interactions by computationally modelling the interactomes of single species such as *Plasmodium falciparum* [Date and Stoeckert, 2006], *Saccharomyces cerevisiae* [Deng et al., 2004, Jansen et al., 2003, Kiemer et al., 2007, Lee et al., 2004], or multi-species such as the STRING database and its accompanying Web interface [von Mering et al., 2007].

Often a large amount of data is available to describe many different ways that proteins can functionally interact. These data often result from high-throughput experimental programmes and can include information about which proteins are co-expressed in a number of microarray studies, for example, through to which protein names are cocited in the abstract of a paper. Therefore, the construction of these networks is often carried out on information extracted from multiple data sources including high-throughput experiments, smaller-scale experiments, curated databases and computational prediction methods. Integrating different forms of evidence, each with varying degrees of reliability and associated biases, effectively provides a more reliable interactome. Furthermore, by combining data sources from studies focusing on different biological properties essentially provides an interactome modelling a broader range of characteristics. When combining data sources, it is important that the relative quality of the data be assessed in order to determine the contribution

of that source to the total edge weight describing the functional association between two proteins [Cusick et al., 2005, Ge et al., 2003, Jansen et al., 2003, Lee et al., 2004]. Probabilistic functional interaction networks (PFINs) describe the functional associations between proteins as a weight on an edge between two genes or gene products in a network. The weight on the edge denotes the probability that the two proteins functionally interact according to some kind of evidence.

However, to be able to integrate data sources that employ their own weighting schemes, or alternatively those that have no weight associated, a unified weighting method needs to be implemented. Most often a *guilt-by-association* approach is adopted, in which a *gold standard* dataset is used from which true positives (TP) and false positives (FP) can be calculated. A commonly used benchmark is the KEGG pathway database, either on its own, as in Lee et al. [2004], or in combination with another benchmark such as Gene Ontology (GO), as in Date and Stoeckert [2006].

However, the use of a gold standard is not the best approach to uniformly weight the data, as it is largely effected by the reliability and content bias of the data source used as the benchmark. In addition, it removes the possibility of incorporating the chosen gold standard dataset as input to the interactome. Methods are therefore required in which integration of data sources do not rely on a gold standard, such as Deng et al. [2004].

In any modelling process, mathematics plays an important role. In terms of interactomes, mathematical methods are required to process the data to construct the interactomes, but methods are also used in analysing the properties of networks; Bayesian statistics are frequently used for this purpose. Bayesian statistics are often applied to biological problems due to their ability to measure the degree of belief in a hypothesis. The Bayesian approach to statistics revolves around the concept that the probability that a hypothesis is true is based on known observations. In other words, the probability of an event occurring is based on the *degree of belief* in that event [Heckerman, 1996].

2.4.2 Interactome comparison

The aim of developing protein interaction networks is to understand how proteins interact and hence how they function, individually and as part of a system. This knowledge can be gained through applying prediction methods within the individual interactomes themselves. The GOM (Global Optimisation Method) [Vazquez et al., 2003] and the improved MFGO (Modified and Faster Global Optimisation) [Sun et al., 2006] methods predict the function of unclassified proteins in the context of the surrounding interaction network. Some methods have been devised that generate subgraphs based on topological statistics, such as TopNet [Yu et al., 2004] and the algorithm termed "local graph alignment" [Berg and Lässig, 2004].

More information can be extracted by making comparison across all the interactomes i.e. across species. A standard approach to annotating a protein with specific functions is to identify proteins with a similar sequence that are well characterised; this involves identifying homologues, or more precisely, orthologues. *Homologs* are proteins that share a common ancestry; this includes *orthologues* that evolve by speciation, or *paralogues* that arise by gene duplication. Orthologs generally retain similar domains and functions following speciation, whereas paralogues evolve new functions. *In-* and *out-paralogues* denote genes duplicated subsequent or prior to speciation, respectively. *Co-orthologues* define orthologues between in-paralogues and genes in the other species [Alexeyenko et al., 2006, O'Brien et al., 2005]. Speciation is defined as the formation of new species that live independently from the species from which they evolved [Hale et al., 1995].

A number of automated methods have been developed for orthologue identification, distinguishing between probable (co-)orthologues and paralogues, including:

- Phylogeny-based methods including:
 - Resampled Inference of Orthology (RIO) [Zmasek and Eddy, 2002],
 - Orthostrapper [Storm and Sonnhammer, 2002].
- Evolutionary distance-based methods including:

- Reciprocal Smallest Distance (RSD) [Deluca et al., 2006, Wall et al., 2003].
- BLAST-based methods including:
 - Reciprocal Best Hit (RBH),
 - COGs [Tatusov et al., 2003],
 - OrthoMCL [Li et al., 2003],
 - Inparanoid [Remm et al., 2001] and Multiparanoid [Alexeyenko et al., 2006].

According to a study carried out in Chen et al. [2007], the best performing algorithms are considered to be OrthoMCL and Inparanoid, which show similar performance when comparing two species. Like Inparanoid, OrthoMCL uses RBH to recognise co-orthologue relationships, then defines orthologue clusters using a Markov Clustering (MCL) algorithm, in which the cluster granularity depends on the *inflation parameter*. OrthoMCL generates clusters of proteins where each cluster consists of orthologues or in-paralogues from at least two species [Li et al., 2003]. A limitation of this method is that it does not provide confidence values for the predicted orthologues. Additionally the orthologue groups do not necessarily have a unique common ancestor, leading to potential inclusion of out-paralogues in the same group. OrthoMCL has the advantage that it is capable of clustering orthologues across multiple species.

The Inparanoid clustering algorithm is limited to pairwise proteome orthology detection, unlike OrthoMCL. However, Inparanoid has the added benefit that it provides confidence scores for the predicted orthologues [Remm et al., 2001]. Multiparanoid extends the Inparanoid algorithm to allow clustering over multiple species, using the output generated by Inparanoid to build multi-species clusters. The drawback of Multiparanoid is that it may only be used when comparing across closely-related species [Alexeyenko et al., 2006].

The performance of Multiparanoid and OrthoMCL are comparable, although differences have been highlighted between these two algorithms. Orthologues identified by OrthoMCL include more out-paralogues compared to Multiparanoid. OrthoMCL

produces tighter clusters reducing tree conflicts, and Multiparanoid has a stricter inclusion policy [Alexeyenko et al., 2006].

Both Inparanoid/Multiparanoid and OrthoMCL are directly applicable to the task of comparing across all the *Bacillus* interactomes to identify *Bacillus*-specific *interologs* i.e. conserved interactions across species [Matthews et al., 2001]. However, since Multiparanoid is only applicable to species that are closely related, this program does not offer the diversity required for future analysis.

The methods described so far focus on comparing interactions individually, from which conserved network regions can be identified (as in Gandhi et al. [2006]), but there are alternative methods that create a global alignment between networks that can also identify conservation. Identifying conserved regions can be used to infer conserved components of the cellular machinery and then use those components to predict new protein functions and interactions. Conserved interactions across species are also less likely to be false positives [Bandyopadhyay et al., 2006]. PathBLAST is a program designed for this purpose, that compares the protein interaction networks of two species to identify conserved network regions, just as BLAST performs pairwise protein sequence alignment [Kelley et al., 2004]. PathBLAST has also been extended to provide a computational framework for comparing multiple protein networks (NetworkBlast) [Sharan et al., 2005]. However, PathBLAST is not freely downloadable, so it cannot be run over the integrated networks of the *Bacillus* species. NetworkBlast bases its initial search on particular types of graphical form modelled on signalling pathways (strings) and protein complexes (clusters). The Græmlin algorithm extended this approach by searching efficiently for similar subgraphs of arbitrary topology across multiple networks [Flannick et al., 2006]. Other methods cluster nodes based on their distance from one another in each network [Ogata et al., 2000], or a combination of the distance of nodes and their edge similarity, as used for comparing metabolic networks in MaWish [Koyutürk et al., 2004]. More recently, a probabilistic model described in Hirsh and Sharan [2007] has been developed to identify conserved protein complexes that describe the evolution of conserved protein complexes from an ancestral species through link dynamics and gene duplication events. This probabilis-

tic model shows comparable performance to NetworkBlast, although NetworkBlast is unable to distinguish conserved motifs to the same degree of detail as this new model. An alternative approach called PHUNKEE (Pairing subgraphs Using NetworkK Environment Equivalence) takes into account gaps in the network structure, allowing for elements that have been inserted or deleted through the course of evolution. The similarity of subgraphs is based on edge, as well as node, similarity. All edges adjoining nodes belonging to a subgraph are considered (the network context), rather than simply the edges connecting nodes within the subgraph [Cootes et al., 2007].

Clustering interactomes to identify regions of highly connected nodes is also an approach that can be employed in analysing a network. Many clustering algorithms have been used to analyse interactomes, including Molecular Complex Detection (MCODE) [Bader and Hogue, 2003], MCL [van Dongen, 2000a], Restricted Neighbourhood Search Clustering (RNSC) [King et al., 2004], Super Paramagnetic Clustering (SPC) [Blatt et al., 1996] and the more recent ensemble framework [Asur et al., 2007]. Clustering is not limited to single interactomes, but can be extended to cross-species analysis, using orthology to infer vertical links between the individual interactomes.

Other approaches have been developed that reverse the idea in identifying conserved interactions, by using conserved interactions to predict functional orthology. It is built on the concept that a protein and its functional orthologue are likely to interact with proteins in their respective networks, that are themselves functional orthologues [Bandyopadhyay et al., 2006].

2.4.3 Interactome visualisation

A variety of software tools are available to visualise and analyse biological networks. To display molecular interaction data as a two-dimensional network, there are general-purpose graph viewers such as Pajek [Batagelj and Mrvar, 1998]. Other viewers link the network to molecular interaction and functional databases such as biopixie for yeast [Myers et al., 2005], IntNet for humans [Xia et al., 2006]. More general viewers for biological networks include Osprey [Breitkreutz et al., 2003] and VisANT [Hu

et al., 2007]. A program called Cytoscape has also been developed; this is a general-purpose open-source modelling environment for integrating biomolecular interaction networks and states [Shannon et al., 2003].

Cytoscape is a modelling environment in which molecules are represented as *nodes* and intermolecular interactions are represented as *edges* between the nodes. A variety of different model parameters describing molecular states and interactions can be integrated into the network as *attributes* on the nodes and edges. Hierarchical data such as ontologies can be incorporated into the network through *annotations*. Using this GUI, selected node and edge attributes and annotations can be viewed.

Cytoscape graphically illustrates a network and its associated integrated data in a variety of graph layouts. It also includes ways of altering the display depending on attribute values. Tools are also provided in Cytoscape for selecting and filtering interesting subsets of nodes and edges. In addition, the core functionality can be extended by using the interface to implement plugin modules to provide new algorithms, additional network analysis, and/or biological semantics [Shannon et al., 2003]; an example is the clustering algorithm MCODE, which is provided as a plugin. The separate modules of which Cytoscape is composed are illustrated in figure 2.14.

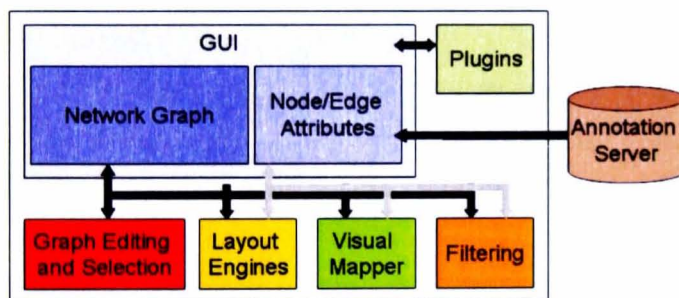


Figure 2.14: The Cytoscape GUI displays the network graph and node/edge attribute data. Graph editing, graph layout, attribute-to-visual mapping, and graph filtering operate over the network and attribute information. Plugins are available that manipulate the network structure in different ways. Annotations are also available through a separate server. Adapted from Shannon et al. [2003].

2.5 Conclusions

The secretome, a subset of the proteome, is a biologically important group of proteins that exert their effect in the extracytoplasmic environment of a cell. For this reason, secretory proteins are of great interest, providing insights into how bacteria interact with their environment, to not only sustain themselves, but in the case of pathogens, to understand how they subvert their hosts. Exposing the functions of secretory proteins can therefore lead to a better understanding of the secretome, and more broadly, the proteome. A number of previous studies have centred on the model Gram-positive bacteria, *B. subtilis* (strain 168) [Antelmann et al., 2001, Tjalsma et al., 2004, 2000]. In some cases these predictions have been experimentally verified.

The extension of this categorisation and analysis process to other Gram-positive bacteria is the next logical step. Comparisons across the different secretomes can then be carried out to reveal potentially interesting differences between species. In fact, the study described in Binnewies et al. [2005] not only predicts secreted proteins, but also performs cross-species comparison between the different secretion systems in a selected number of Gram-positive and Gram-negative bacteria to identify conserved sequences.

Furthermore, analysing proteins in the context of their interactions can aid in the understanding of cellular function, and the possible function of uncharacterised proteins. Proteins interact in myriad ways within cells, forming a complex network of structural and functional processes. Several methods of constructing PFINs have been developed, all of which integrate information from a variety of sources. The proteomes of eukaryotic species have largely been the focus of attention; the most intensely studied being the model organism *S. cerevisiae* [Deng et al., 2004, Jansen et al., 2003, Kiemer et al., 2007, Lee et al., 2004].

To date there have been relatively few reports of the application of PFINs to prokaryotes. Interactomes representing physical interactions in the entire bacterial proteome of *E. coli* [Butland et al., 2005] and *H. pylori* [Rain et al., 2001], have been developed [Noirot and Noirot-Gros, 2004]. In addition, models of networks represent-

ing specific biological processes have also been constructed, including DNA replication in *B. subtilis* [Noirot-Gros et al., 2002] and type IV secretion systems in *Agrobacterium tumefaciens* [Ward et al., 2002] and *Rickettsia sibirica* [Malek et al., 2004]. The construction and analysis of these networks has led to the annotation of a number of previously uncharacterised proteins. In the case of Rain et al. [2001], *H. pylori* was compared to *E. coli*, from which knowledge about proteins of unknown function in *H. pylori* were identified through homologue detection and function assignment based on the *E. coli* interaction network.

In addition, it is also valuable to compare interactomes across species. Cross-species interactome comparisons enable similarities and dissimilarities to be observed, which can then be used to generate hypotheses about issues such as evolutionary pathways or missing interactions. Many methods have been developed for this purpose, facilitating the process of cross-species interactome comparison.

The application of such methods has previously been shown by comparing eukaryotes, for instance yeast, worm and fly [Sharan et al., 2005], and also with human [Gandhi et al., 2006]. The majority of studies focus on these model organisms, with comparisons between prokaryotes being few and far between. However, a study has previously made comparisons between a eukaryote and prokaryote, *S. cerevisiae* and *Helicobacter pylori* [Kelley et al., 2003], resulting in the prediction of a functionally unknown protein in *H. pylori* as a DNA polymerase, and another as a membrane-specific protein. The latter protein was found to belong to a pathway that shares homology with the yeast nuclear pore complex. Another includes the prokaryotic-specific comparisons between *H. pylori* and *E. coli* [Rain et al., 2001]. Through these studies, cross-species specific information can be extracted. The value of interactome comparison techniques has therefore been recognised, and is an area of active research. There is however still much scope for improvement, particularly for prokaryotes.

Chapter 3

An e-science approach to the genomic scale characterisation of bacterial secreted proteins

3.1 Introduction

The aim of this study was to identify proteins that are likely to be secreted from the cytoplasm and to classify them according to the putative mechanism of their secretion. Such proteins include those exported to the extracellular medium, as well as proteins that attach themselves to the outer surface of the membrane (lipoproteins) and cell wall binding proteins (sortase mediated proteins containing an LPXTG motif). Following the classification of the secreted proteins, the composition of the predicted secretomes in the 12 species under study were then investigated.

The BaSPP (Bacterial Secretory Protein Prediction) package was designed and implemented in order to meet this objective. The central components to this system are two Web service-based SCUFL workflows: the *classification workflow* and the *analysis workflow*. The classification workflow is concerned with making predictions about the secretory characteristics of a particular protein from a given set of proteins. The analysis workflow processes the data from the first workflow in order to analyse the function of the secreted proteins that have been identified.¹

¹Presented at UK e-Science All Hands Meeting 2006 [Craddock et al., 2006].

3.1.1 Computational approaches to predicting and classifying secreted proteins

Computational approaches are already available to aid in classifying proteins according to their secretion potential. These tools generally look for a pattern consistent with a specific type of protein, by often implementing a learning algorithm such as Hidden Markov Model (HMM) and neural networks. The most relevant programs used to characterise secreted proteins apply one or both of these learning algorithms, including:

- **SignalP**

SignalP² predicts the presence of signal peptides and the SPase I cleavage site. It uses both neural networks and HMMs [Nielsen and Krogh, 1998].

- **LipoP**

LipoP³ predicts lipoproteins in Gram-negative bacteria [Juncker et al., 2003]. Despite the specificity of LipoP for Gram-negative lipoprotein signal peptides, the HMM algorithm was shown to be able to identify a high proportion of Gram-positive lipoproteins. The HMM algorithm differentiates between lipoproteins (SPase II cleaved proteins), SPase I cleaved proteins, cytoplasmic proteins and transmembrane proteins [Juncker et al., 2003].

- **TMHMM, MEMSAT and TMAP**

An evaluation of the tools available to predict the topology of transmembrane proteins is detailed in Möller et al. [2001]. From this study, it was determined that TMHMM 2.0 is the most reliable and accurate. The approach adopted in TMHMM⁴ uses an HMM algorithm to determine the most probable topology for the entire protein [Krogh et al., 2001, Möller et al., 2001].

²SignalP Website: <http://www.cbs.dtu.dk/services/SignalP/>

³LipoP Website: <http://www.cbs.dtu.dk/services/LipoP/>

⁴TMHMM Website: <http://www.cbs.dtu.dk/services/TMHMM/>

The runners up to TMHMM included MEMSAT and TMAP. MEMSAT and TMAP implement a combined approach in which global heuristics operate over local level results [Möller et al., 2001]. MEMSAT⁵ uses well-characterised membrane protein data (from bacteria, plants and animals) to calculate loglikelihoods with the aim of identifying biases towards certain amino acids. A novel dynamic programming algorithm is then used to predict the membrane structure by expectation maximisation [Jones et al., 1994]. TMAP⁶ uses a prediction algorithm based on multiple sequence alignments of related proteins [Persson and Argos, 1994].

- **ps_scan**

The standalone version of ScanProsite⁷, known as ps_scan, is a Perl program that provides a means to scan one or several PROSITE patterns, rules and profiles against one or several protein sequences in SWISSPROT or FASTA format [Gattiker et al., 2002].

3.2 Results

3.2.1 System architecture

The architectural view of the workflow components involved in the prediction and analysis of the secretomes of the *Bacillus* species is shown in figure 3.1.

A general feature of the workflows is their linear construction. In the classification workflow for example, at each step the set is reduced in size, removing those proteins that do not require further classification. The results of the classification process are stored in a remote relational database and the reduced set of proteins passed to the next service in the workflow. In the analysis workflow, data derived from the classification workflow is retrieved from the database and analysed sequentially using a further series of service enabled tools. Compare these linear workflows to the

⁵MEMSAT Website: <http://saier-144-37.ucsd.edu/memsat.html>

⁶TMAP Website: <http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+lib+TMAP>

⁷ScanProsite Website: <http://expasy.org/tools/scanprosite/>

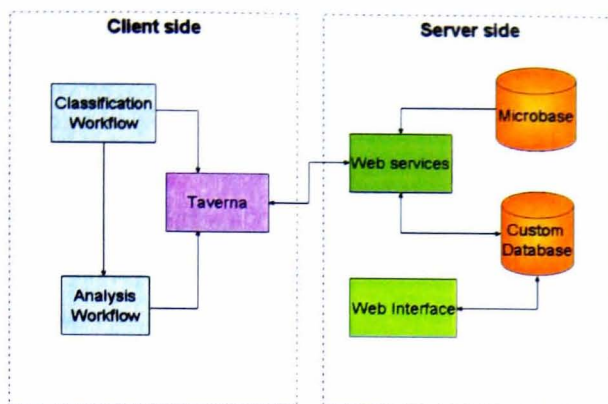


Figure 3.1: Architectural layout of the classification and analysis workflow components, illustrating the client side workflow execution using Taverna, and the hosting of Microbase, the custom database, Web services and the Website on the server side.

fanned/tree-like Graves' Disease workflow [Li et al., 2004], which instead performs all tasks in-parallel.

Web services were not implemented on the client machine that enacted the workflow, but were located as close to the database as possible to maximise performance. This is particularly important for those services interacting with the database, to maximise the efficiency of transfer of large genomic and proteomic datasets from the service to the database. The services were orchestrated using SCUFL workflows and constructed and enacted using the Taverna workbench [Oinn et al., 2004] .

Most of the services used within the workflows both use and return text-based information in standard bioinformatics formats, such as FASTA format. At all steps of the classification workflow, the intermediate results were extracted into a custom-designed database. This was achieved using a set of bespoke data storage services which parse the raw textual results and store them in a structured form. It is this structured data that is used to feed later services in both the classification and analysis workflows. The custom database was hosted on a machine in close network proximity to the Web services. This has the significant advantage of reducing the network costs involved in transferring data to and from the database.

After completion of the classification workflow, the custom database contains the data relating to each protein analysed, including the raw data as well as an integrated summary of the analysis. Tracking the provenance of the data is important in this context because there are a number of different routes for the classification workflow to designate a protein as secretory. The basic operational provenance provided by Taverna also aided in the identification of service failures [Zhao et al., 2004]. This was particularly important while running the transmembrane domain prediction services, as these run concurrently; a failure in one, therefore, does not impact on the execution of the classification workflow. However, since the data returned may be incomplete a failure needs to be recorded.

A Web portal was developed to provide a user-friendly and familiar mechanism for accessing the secretomes data in the database⁸. From this site, users can select the bacterial species in which they are most interested and view the corresponding results. Data is initially displayed as a table, documenting the entire proteome for the selected bacteria in terms of gene names, locus tags, gene location, Uniprot reference and classification workflow prediction. The colour coding on the left hand side of this table indicates the set of parameters used to generate the results. This allows for multiple workflows to be run on the same bacterium using different parameter sets. From the results page, the users can navigate further to view the details of the classifications. The protein sequences may be viewed along with an overlay of predicted signal peptides and their cleavage sites. Users may also edit and curate the database by adding comments on the findings of the classification workflow as appropriate. A screenshot of the database portal is shown in figure 3.2.

In both the classification and analysis workflows, all legacy programs were implemented locally, but exposed using Web services.

⁸BaSPP Website: <http://bioinf.ncl.ac.uk:8081/BaSPP>

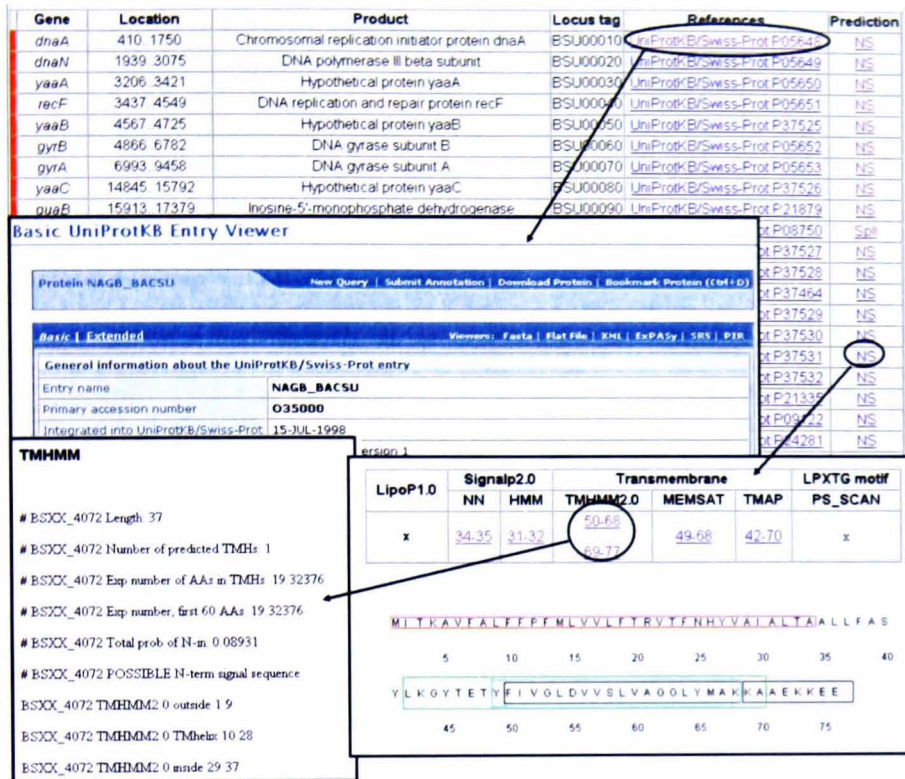


Figure 3.2: Screenshot of the Web portal summarising the characteristics of predicted secreted proteins from *B. subtilis* (strain 168).

3.2.2 Development of a secretory protein classification workflow

The first objective of the classification workflow was the prediction of lipoproteins (figure 3.3). This was the responsibility of the first service in the workflow, employing LipoP 1.0 [Juncker et al., 2003], which takes as input a set of all predicted proteins derived from the EMBL record for the complete genome sequence. Proteins are predicted to be lipoproteins if the LipoP results predict an SPase II cleaved signal peptide above the threshold cutoff score of 3.0. The putative type II signal peptides (SpII) are then checked for transmembrane domains, whereas the non-lipoproteins are checked for the presence of an SPase I signal peptide.

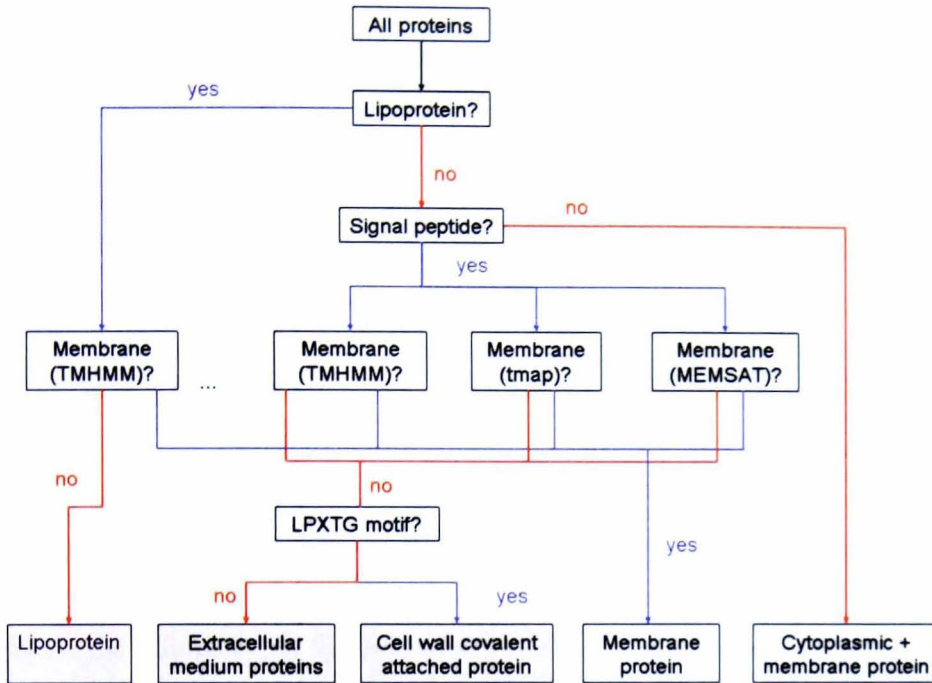


Figure 3.3: Basic representation of the functionality of the classification workflow. Shaded boxes indicate the set of secreted proteins.

The prediction of the presence of an SPase I signal peptide and the location of the cleavage site is accomplished by using a SignalP 2.0 Web service [Nielsen and Krogh, 1998]. SPase I prediction was performed after the lipoprotein identification because of possible limitations in the efficiency of SignalP at detecting lipoproteins. The use of SignalP at this point also removes most of the proteins from subsequent analysis. Reducing the size of the dataset has considerable advantages as the downstream analyses are potentially computationally intensive.

SignalP implements two methods to provide separate prediction methods that utilise neural networks and HMMs. A combination of both these methods are used to predict whether a protein is a specific substrate for SPase I. A positive result from both neural networks and HMM is required in order to predict a protein to contain a

type I signal peptide (SpI).

Proteins with SpI-like or SpII-like sequences, but which have additional transmembrane domains, are likely to be retained in the membrane. To identify these putative membrane proteins from among the proteins in either the SpI or SpII datasets, a combination of three transmembrane prediction Web services were used, based on the tools TMHMM, MEMSAT and TMAP, respectively [Möller et al., 2001]. A subsequent service in the workflow was responsible for integrating the results derived from these three tools, to make a final prediction about the presence of a putative transmembrane protein. The prediction of a transmembrane protein was determined if any one of the tools made a positive prediction. Each protein, whether a SpI or SpII protein, had their first 40 amino acid residues removed and the remaining sequence was passed on for analysis by the various transmembrane tools. This cleavage step was to ensure the removal of the hydrophobic domain within the signal peptide region, which would ultimately be detected as a transmembrane domain. Cleavage of the first 30 or 35 amino acids may not result in the complete removal of this hydrophobic domain.

In *theory*, using a combination of programs to make predictions in this way should provide more meaningful results. However, based on selected manual curation of results provided by each of the three transmembrane prediction tools, it was decided to rely solely on the prediction of TMHMM 2.0, as this tool provided the most accurate prediction.

The subset of the lipoproteins without a predicted transmembrane domain were categorised as being lipoproteins and were not subject to further analysis. However, the subset of SpI proteins corresponding to proteins with no predicted transmembrane domain were further analysed for the presence of the cell wall binding amino acid motif, LPXTG using the tool, ps_scan that was wrapped as a Web service and called from the workflow.

3.2.2.1 Workflow design and implementation

Closer inspection of the classification workflow highlights the complexity involved in the categorisation process. For simplicity, the construction of this workflow was divided into a number of subworkflows (figure 3.4).

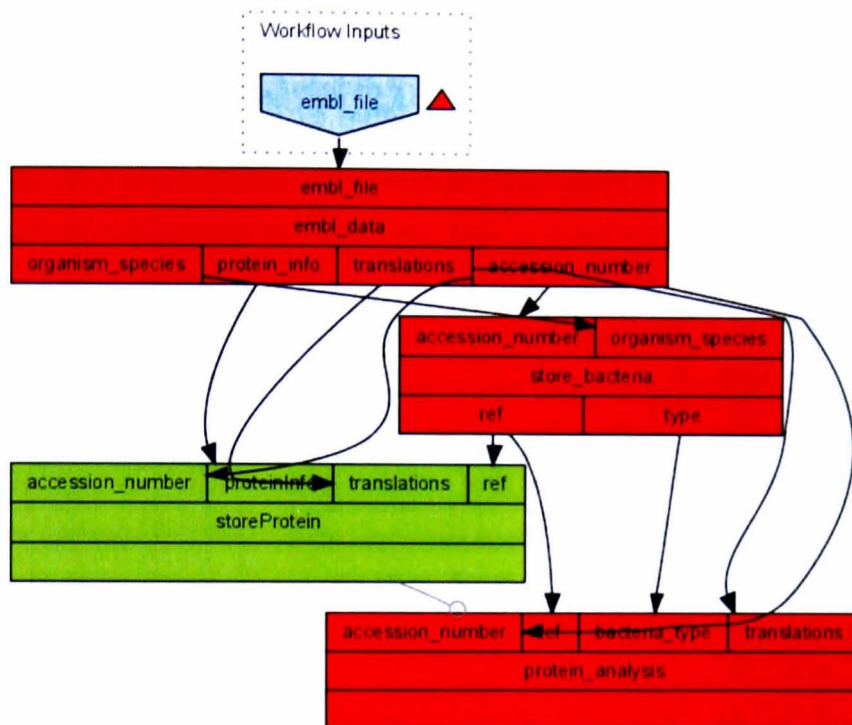


Figure 3.4: Taverna representation of the classification workflow consisting of three subworkflows (red) (*embl_data* extracts information from the EMBL file, *store_bacteria* stores information specific to the bacteria, *protein_analysis* analyses the proteome) and one Web service (green) (*storeProtein* which stores information specific to the proteome).

The workflow is initiated by providing an EMBL file as input, that annotates a particular bacterial genome. EMBL files provide information in a specific file format in order to be human as well as computationally readable. A number of specialised Web services within the classification workflow are dedicated to processing the information

contained within this file. These services use BioJava⁹ to extract the information about an organism's genome and its associated annotated features (figure 3.5). Those features that were extracted that relate to the identity of the organism itself include accession number (AC) and organism species (OS). The annotated features of interest within this study are the coding regions (CDS). The information relating to these CDS features were extracted from the EMBL file, include translation, dbxref, gene, locus tag, protein_id, product and note (table 3.1).

Overall, a total of $\sim 57,000$ proteins were extracted from all 12 genomes supplied as input. The data is stored in the database, but only if the organism supplied as input to the workflow has not previously been analysed using the same parameter settings for the different programs. If the EMBL data has already been stored in the database after a previous workflow execution using the same parameters, the workflow fails, preventing duplicate entries in the database (figure 3.6 and table 3.2).

A number of bioinformatics tools, previously highlighted in section 3.1.1, have been wrapped as Web services in order to categorise secreted proteins. Additional services are provided which store the data generated by each of these tools in a custom database (figure 3.7 and table 3.3). Furthermore, filter services allow proteins of interest to be provided as input to the services downstream. The filtering process is tracked in the database. Essentially the filtering reduces the data being processed at each phase in the workflow (shown in figure 4.1). Starting with $\sim 57,000$ proteins to the first service, this number is reduced as the proteins proceed through the workflow until all the proteins have been classified.

Error handling can be incorporated into the workflow using an additional feature of Taverna which allows alternative services to be executed, allowing the database to be cleared of all entries which may have been added during a failed workflow run.

As a number of studies have previously been carried out to identify components of the *B. subtilis* secretome, comparing the classification workflow results for this organism to the documented findings, provides a means to validate the results generated through the characterisation process described.

⁹BioJava Website: <http://biojava.org/wiki/>

Service	Input	Output	Description
getAccessionNumber	EMBL file	accession number	Unique identifier for each EMBL entry.
getOrganismSpecies	EMBL file	organism species	Preferred scientific name of the organism in EMBL entry.
getTranslations	EMBL file	translation	Amino acid sequence resulting from the sequence of nucleotides
getProteinId	EMBL file	list of protein information:	
		dbxref	Database cross-references.
		gene	Gene symbol.
		locus tag	Unique tag.
		protein_id	Protein identifier.
		product	Product name.
		note	Any additional information.

Table 3.1: The Web services used in the classification subworkflow, *embl_data*. All take as input an EMBL file (in String format).

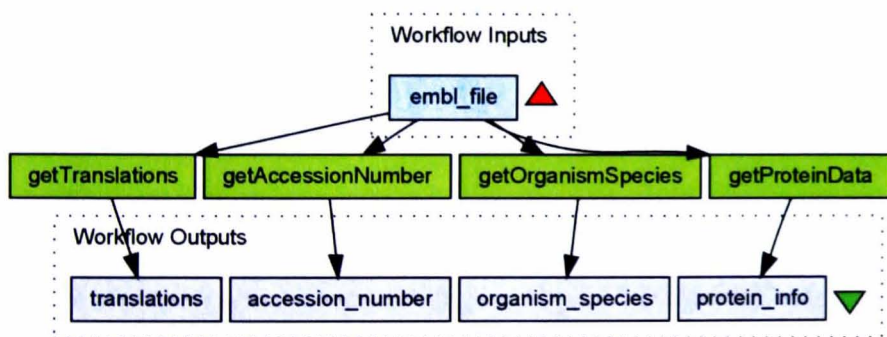


Figure 3.5: Taverna representation of the classification subworkflow, *embl_data*. It consists of four Web services (green), each responsible for extracting different data from an EMBL file.

Service	Input	Output	Description
checkParameters	accession number	accession number	Fails if bacteria has been analysed with same parameters already. (Parameters are set in database before running the workflow.)
storeBacteria	accession number	accession number	Stores bacterial information.
storeParameters	accession number, type of bacteria (gram+/gram-)	parameter reference id	Associates set parameters with bacteria being analysed.

Table 3.2: The Web services used in the classification subworkflow, *store_bacteria*.

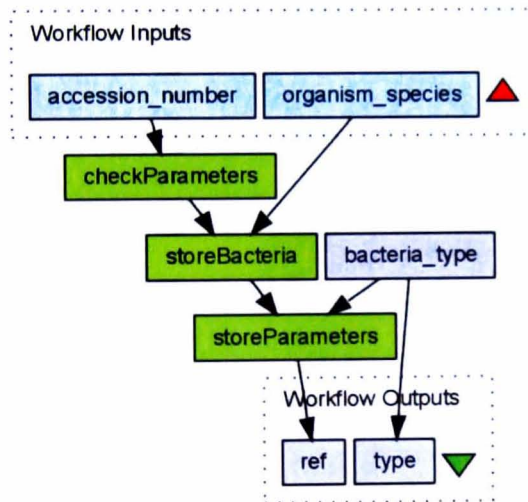


Figure 3.6: Taverna representation of the classification subworkflow, *store_bacteria*, consisting of three Web services (green). If the workflow has not previously analysed this bacteria using the same program parameters i.e. there is no current entry in the database (determined using *checkParameters*), then the data identifying the bacterial species is stored in the database (using *storeBacteria* service), as are the program parameters (using *storeParameters* service).

Service	Input	Output	Description
PROGRAMS (TAVERNA ITERATION PROCEDURE EXECUTES LISTS)			
lipop	protein	LipoP results	Predicts lipoproteins.
signalp	protein, parameters	SignalP results	Predicts signal peptides.
tmhmm memsat tmap	protein (in case of MEMSAT, parameters too)	Program results	Predicts transmembrane domains.
ps_scan	protein, motif id	ps_scan results	Identifies motifs.
PROGRAM PARAMETERS (FROM DATABASE)			
getMethodParameter getTruncateParameter		method (nn/hmm) truncation value	SignalP parameters.
getMinimumLoopLength getHelixScoreCutoff getMaxLengthHelix getMinSeqLength getMinLengthHelix getMaxHelices		min. loop length helix score cutoff max length helix min. sequence length min. length helix max. helices	MEMSAT parameters.
STORE TO DATABASE			
store_lipop_results store_signalp_results store_tmhmm_results store_memsat_results store_tmap_results store_psscan_results	list of respective program results		Stores results in database.
FILTER RESULTS			
filterLipopResults filterSignalpResults filterTMHMMResults filterMemsatResults filterTmapResults	accession number, parameter set id	protein lists	Gets appropriate results from database for further analysis.
getLipoproteins getNonLipoproteins	accession number, parameter set id	protein lists	Gets lipoproteins/ non-lipoproteins.
BEANSHELLS			
get_input_as_list	protein list	protein list	Reformats input.
merge_lists	lipoproteins list, signal peptides list	combined list	Merges lists.
append_tags_4_tmhmm append_tags_4_memsat	lipoproteins list and signal peptides list,	tagged lipoproteins list and signal	Appends tags to distinguish between
append_tags_4_tmap	in a list (list of lists)	peptides list, in a list (list of lists)	lipoproteins and signal peptides.

Table 3.3: The Web services used in the classification subworkflow, *protein_analysis*.

3.2.2.2 Comparison of BaSPP-based predictions for *B. subtilis* secreted protein to experimentally verified protein annotations

The classification workflow predictions for *B. subtilis* were compared to a set of 47 proteins of *B. subtilis* whose secretory status has been experimentally verified and documented in the literature. Of these 47 proteins:

- 18 out of 19 were correctly identified as SpI (table 3.4);
- 11 of 11 were correctly identified as SpII (table 3.5);
- all 11 transmembrane proteins and 7 cytoplasmic proteins were correctly predicted not to be secreted (tables 3.6 and 3.7).

One SpI specific protein, PhoD, was not detected. The unusually long signal peptide sequence of PhoD, 51 amino acid residues according to Tjalsma et al. [2000], may explain why SignalP does not predict this protein to be targeted by SPase I. Despite the specificity for SPase I, PhoD is actually transported by the Tat pathway, not Sec. As detailed in Bendtsen et al. [2005], PhoD is not predicted to be secreted by SignalP, but it is positively identified using TatP¹⁰ and TATFIND¹¹.

Four transmembrane proteins (encoded by *fruA*, *pyrP*, *rbsC* and *tagH*) were also not identified by the classification workflow. This omission was because these proteins had been removed from the classification process, due to the lack of a positive SpI or SpII prediction, before TMHMM analysis. The important point is none of these known transmembrane proteins were predicted as being secreted. The intention of the classification workflow was only to classify secreted proteins, not cytoplasmic and transmembrane proteins. If this were the intention, the workflow would have been designed differently, in which every protein in the proteome would be analysed using TMHMM. TMHMM analysis of every protein would have resulted in the positive identification of these four proteins as transmembrane.

¹⁰TatP Website: <http://www.cbs.dtu.dk/services/TatP/>

¹¹TATFIND Website: <http://signalfind.org/tatfind.html>

Gene	Description	Locus tag	Prediction	Reference
<i>abnA</i>	arabinan-endo 1,5-alpha-L-arabinase	BSU28810	SpI	[Leal and de Sá-Nogueira, 2004]
<i>amyE</i>	alpha-amylase	BSU03040	SpI	[Hirose et al., 2000]
<i>bglS</i>	endo-beta-1,3-1,4 glu- canase	BSU39070	SpI	[Murphy et al., 1984]
<i>bpr</i>	bacillopeptidase F	BSU15300	SpI	[Sloma et al., 1990, Wu et al., 1990]
<i>cotN</i>	translocation-dependent antimicrobial spore component	BSU24620	SpI	[Hirose et al., 2000]
<i>csn</i>	chitosanase	BSU26890	SpI	[Hirose et al., 2000]
<i>epr</i>	extracellular serine protease	BSU38400	SpI	[Sloma et al., 1988]
<i>ggt</i>	gamma-glutamyltranspeptidase	BSU18410	SpI	[Xu and Strauch, 1996]
<i>glpQ</i>	glycerophosphoryl diester phosphodiesterase	BSU02130	SpI	[Antelmann et al., 2000]
<i>mpr</i>	extracellular metalloprotease	BSU02240	SpI	[Rufo et al., 1990]
<i>pel</i>	pectate lyase	BSU07560	SpI	[Hirose et al., 2000]
<i>penP</i>	beta-lactamase precursor	BSU18800	SpI	[Hirose et al., 2000]
<i>phoA</i>	alkaline phosphatase A	BSU09410	SpI	[Hulett et al., 1991]
<i>phoB</i>	alkaline phosphatase III	BSU05740	SpI	[Antelmann et al., 2000, Hulett et al., 1991]
<i>phoD</i>	phosphodiesterase/alkaline phosphatase D	BSU02620	-	[Eder et al., 1996]
<i>vpr</i>	extracellular serine protease	BSU38090	SpI	[Hirose et al., 2000]
<i>wapA</i>	cell wall-associated protein precursor	BSU39230	SpI	[Hirose et al., 2000]
<i>wprA</i>	cell wall-associated protein precursor	BSU10770	SpI	[Hirose et al., 2000]
<i>xynA</i>	endo-1,4-beta-xylanase	BSU18840	SpI	[Roncero, 1983]

Table 3.4: Comparison of BaSPP SpI protein predictions to a list of experimentally verified SpI proteins. (*phoD* was not predicted to encode a secreted protein by BaSPP as no lipoprotein or signal peptide was detected.)

Gene	Description	Locus tag	Prediction	Reference
<i>appA</i>	oligopeptide ABC transporter (oligopeptide-binding protein)	BSU11380	SpII	[Koide and Hoch, 1994]
<i>araN</i>	arabinose ABC transporter (arabinose-binding protein)	BSU28750	SpII	[Sá-Nogueira et al., 1997]
<i>dppE</i>	dipeptide ABC transporter (dipeptide-binding protein)	BSU12960	SpII	[Mathiopoulos et al., 1991]
<i>glnH</i>	glutamine ABC transporter (glutamine-binding protein)	BSU27440	SpII	[Wu and Welker, 1991]
<i>mntA</i>	manganese ABC transporter (membrane protein)	BSU30770	SpII	[Bartsevich and Pakrasi, 1995]
<i>oppA</i>	oligopeptide ABC transporter (oligopeptide-binding protein)	BSU11430	SpII	[Perego et al., 1991]
<i>opuBC</i>	choline ABC transporter (choline-binding protein)	BSU33710	SpII	[Kappes et al., 1999]
<i>opuCC</i>	glycine betaine/carnitine/ choline ABC transporter (osmoprotectant-binding protein)	BSU33810	SpII	[Kappes et al., 1999]
<i>prsA</i>	molecular chaperone	BSU09950	SpII	[Kontinen et al., 1991]
<i>pstS</i>	phosphate ABC transporter (phosphate-binding protein)	BSU24990	SpII	[Allenby et al., 2004]
<i>rbsB</i>	ribose ABC transporter (ribose-binding protein)	BSU35960	SpII	[Woodson and Devine, 1994]

Table 3.5: Comparison of BaSPP SpII protein predictions to a list of experimentally verified SpII proteins.

Gene	Description	Locus tag	Prediction	Reference
<i>cydD</i>	ABC membrane transporter (ATP-binding protein)	BSU38730	TM	[Winstedt et al., 1998]
<i>dltB</i>	D-alanine esterification of lipoteichoic acid and wall teichoic acid (D-alanyl transfer from Dcp to undecaprenol-phosphate)	BSU38510	TM	[Perego et al., 1995]
<i>fruA</i>	phosphotransferase system (PTS) fructose-specific enzyme IIABC component	BSU14400	-	[Reizer et al., 1999]
<i>ftsH</i>	cell-division protein and general stress protein (class III heat-shock)	BSU00690	TM	[Deuerling et al., 1997]
<i>ftsX</i>	cell-division protein	BSU35250	TM	[de Leeuw et al., 1999]
<i>fcpB</i>	methyl-accepting chemotaxis protein	BSU31260	TM	[Hanlon and Ordal, 1994]
<i>pyrP</i>	uracil permease	BSU15480	-	[Turner et al., 1994]
<i>rbsC</i>	ribose ABC transporter (permease)	BSU35950	-	[Park and Park, 1999]
<i>secDF</i>	protein-export membrane protein	BSU27650	TM	[Bolhuis et al., 1998]
<i>secY</i>	preprotein translocase subunit	BSU01360	TM	[Suh et al., 1990]
<i>tagH</i>	ATP-binding protein	BSU35700	-	[Lazarevic and Karamata, 1995]

Table 3.6: Comparison of BaSPP transmembrane predictions to a list of experimentally verified transmembrane proteins. (Those not precisely identified by TMHMM as transmembrane were not predicted by LipoP or SignalP to be secreted, hence they were not analysed by TMHMM for transmembrane domains.)

Gene	Description	Locus tag	Prediction	Reference
<i>eno</i>	enolase	BSU33900	-	[Leyva-Vazquez and Setlow, 1994]
<i>groEL</i>	class I heat-shock protein (chaperonin)	BSU06030	-	[Schumann et al., 1998]
<i>kata</i>	vegetative catalase 1	BSU08820	-	[Hirose et al., 2000]
<i>pdhD</i>	dihydrolipoamide dehydrogenase E3 subunit of both pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase complexes	BSU14610	-	[Borges et al., 1990]
<i>rocA</i>	pyrroline-5 carboxylate dehydrogenase	BSU37780	-	[Calogero et al., 1994]
<i>rocF</i>	arginase	BSU40320	-	[Gardan et al., 1995]
<i>sodA</i>	superoxide dismutase	BSU25020	-	[Hirose et al., 2000]

Table 3.7: Comparison of BaSPP cytoplasmic protein predictions to a list of experimentally verified cytoplasmic proteins.

3.2.2.3 Comparison of BaSPP-based predictions for *B. subtilis* secreted protein to previously published, computationally generated, annotations

The putative *B. subtilis* secretome was also compared to a study in which the secretome of *B. subtilis* has been computationally determined and, to some extent, experimentally confirmed by proteomic studies [Tjalsma et al., 2004, 2000]. The vendrogram shown in figures 3.8 and 3.9 summarise how the predictions from this study and that of Tjalsma and co-workers compare.

In order to compare the data shown in Tjalsma et al. [2000] with BaSPP more easily, the BaSPP-predicted secretome of *B. subtilis* was formatted in the same way as Tjalsma et al. [2000] (shown in appendix A). A detailed comparison of results are highlighted over the formatted data shown in tables A.1 and A.2, representing the putative SpI and SpII proteins respectively. The 22 SpI proteins and 15 SpII proteins identified by Tjalsma et al. [2000], but not by the classification workflow, are shown in tables A.3 and A.4 respectively.

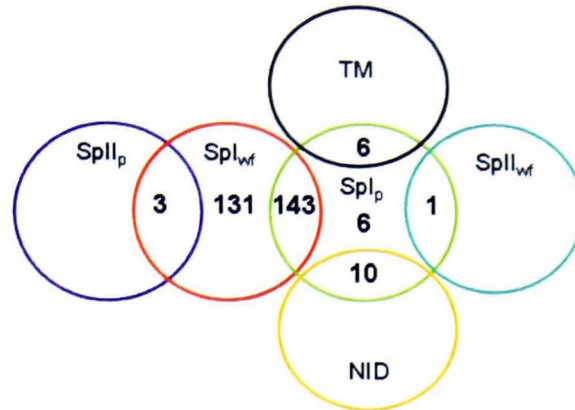


Figure 3.8: Comparison of SpI predictions made by the workflow (Wf) and those described by Tjalsma et al. [2000] (P). The majority of SpI proteins identified in the literature are also identified by the workflow. The remaining SpI proteins described in the literature are either categorised as transmembrane proteins (TM) or SpII proteins by the workflow, or are not found in the dataset supplied as input to the workflow (NID).

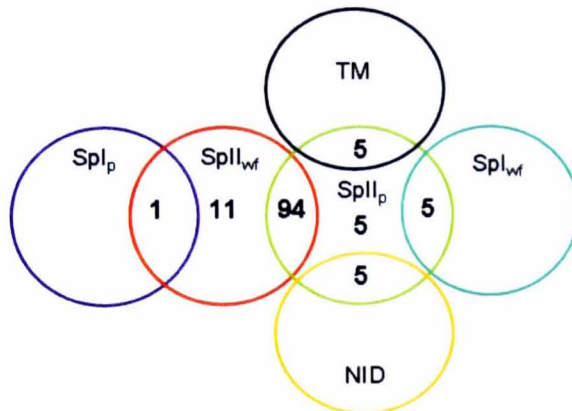


Figure 3.9: Comparison of SpII predictions made by the workflow (Wf) and those described by Tjalsma et al. [2000] (P). The majority of SpII proteins identified in the literature are also identified by the workflow. The remaining SpII proteins described in the literature are either categorised as transmembrane proteins (TM) or SpI proteins by the workflow, or are not found in the dataset supplied as input to the workflow (NID).

3.2.3 Protein analysis workflow

The analysis workflow was designed to process the data about the secretory proteins derived from the classification workflow to provide information about the relationships between the secretome composition of the 12 different organisms in the study. Putative secreted proteins were extracted from the database, clustered into families based on sequence similarity, and the structure, functional composition and relationships between these families were studied (figure 3.10). The set of secreted proteins includes those predicted to be lipoproteins, cell wall binding or extracellular. Transmembrane proteins and cytoplasmic proteins were disregarded.

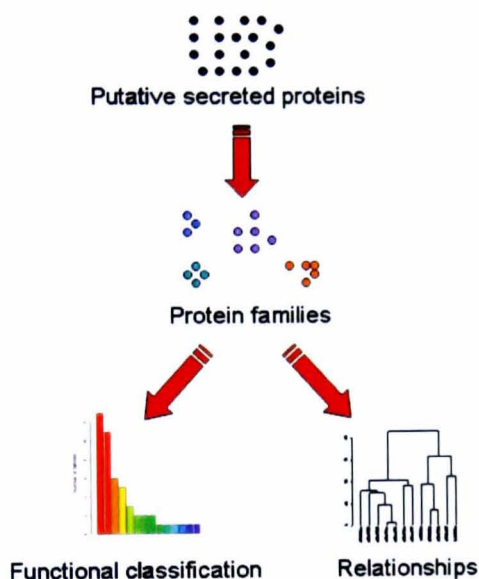


Figure 3.10: Cartoon representation of the functionality of the analysis workflow. All the putative secreted proteins (identified by the classification workflow), for all 12 *Bacillus* strains, are clustered into protein families, from which functional properties of the secretome, and relationships between the secretomes of different organisms, can be determined.

The analysis workflow therefore follows on from the classification workflow, implementing the following procedure (figure 3.11). Analysis of the data was initiated by clustering the putative secretory proteins into protein families. In order to per-

form clustering, the close relatives of the predicted secreted proteins were identified using BLASTp, as the BLASTp algorithm provides a score for the putative similarity between proteins. The necessary BLASTp data was retrieved from the Microbase system, a Web service enabled resource for the comparative genomics of microorganisms [Sun et al., 2005].

Using the BLASTp data, protein families were identified using the MCL algorithm. MCL provides a computationally efficient means of clustering large datasets. A single option, the inflation value, controls the granularity of the output clustering. It is usually chosen somewhere in the range [1.2-5.0]. A value of 5.0 will tend to result in fine-grained clusterings whereas a value of 1.2 will tend to result in very coarse grained clustering. The value used depends on the characteristics of the data to be clustered [van Dongen, 2000b].

This approach of combining BLASTp data with MCL follows that documented in Enright et al. [2002]. For each predicted secreted protein, similar proteins with a BLASTp expect value of less than $1e^{-10}$ were used as input to MCL version 05-321 (inflation value 3.0), providing a cutoff score would remove many of the protein families containing a single protein [Enright et al., 2002][van Dongen, <http://micans.org/mcl/>].

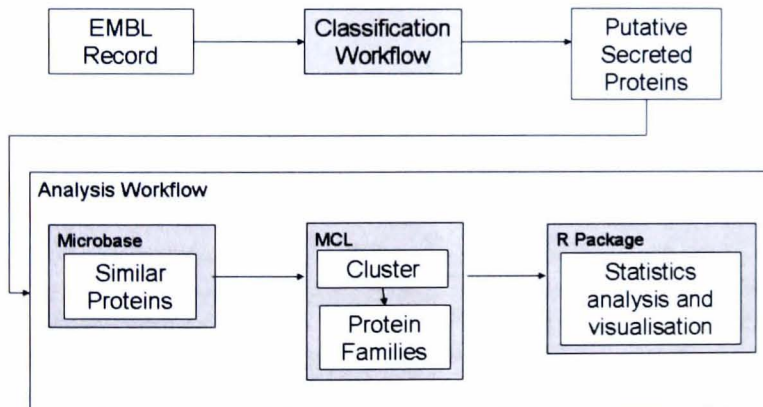


Figure 3.11: Data flow through the classification and analysis workflows.

Hierarchical clustering was performed to identify phylogenetic relations between the *Bacillus* species in the context of their contributions to the secreted protein families. The R package¹² was wrapped as a Web service for this purpose. R is a package providing a statistical computing and graphics environment. A distance matrix was constructed using the Euclidean distance, and clustering was carried out using a complete linkage method.

3.2.3.1 Workflow design and implementation

Unlike the classification workflow, the Taverna representation of the analysis workflow is far more simple (figure 3.12). The bioinformatics tools used were, again, wrapped as Web services, using additional 'shim' services to parse and store the input and output data generated by each of these tools in a custom database.

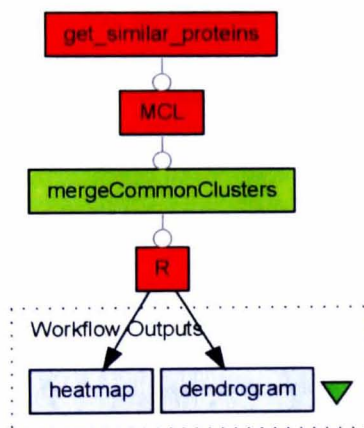


Figure 3.12: Taverna representation of the analysis workflow including two subworkflow (red) (*get_similar_proteins* which extracts the classified secreted proteins from the database and gets similar proteins from Microbase, *MCL* performs clustering on this BLAST weighted protein dataset, and *R* performs statistical analysis on the clustered families) and one Web service (green) (*mergeCommonClusters* which merges common families that were initially separate after performing MCL).

¹²R Website: <http://www.r-project.org/>

Due to the fact the time required for the services *get_similar_proteins* and *MCL* to complete, exceeded the timeout in Taverna, a service failure was assumed by Taverna for both of these services. Therefore, an additional frontend to these services was implemented that utilises threads on the server side to run the service and update the status on completion. On the client side, a feature of Taverna is used that implements a specified time delay to re-run a service on failure. The subworkflow (figure 3.13) initially runs the job, then simply checks the status of the job on the server side at regular intervals. While the job is still running, the status service is designed to fail, which after a certain time delay is then re-executed to again check the status of the job. On completion of the server side program, the status is updated, and consequently Taverna registers a completed service execution. The following service in the subworkflow then retrieves the results.

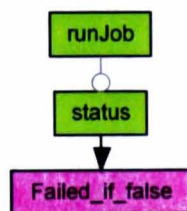


Figure 3.13: Taverna representation of the standard design of the subworkflows *get_similar_proteins* and *MCL*. It checks the status of a thread on the server to determine when a job has finished executing.

For the R subworkflow, firstly a matrix is constructed (*Bacillus* species vs. protein families). This matrix is used by R to produce the dendrogram and heatmap diagrams used to represent the relations between the bacteria graphically, based on the proteins identified in their secretome (figure 3.14).

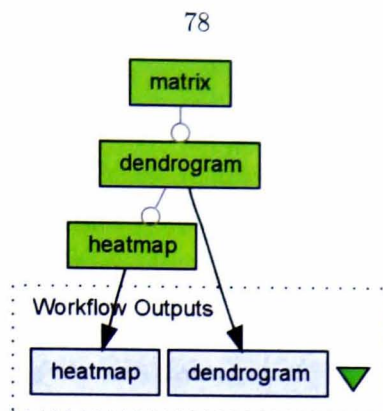


Figure 3.14: Taverna representation of the *R* subworkflow used to generate the dendrogram and heatmaps.

3.2.3.2 Comparison of *Bacillus* secretory protein families to COGs

The MCL output was verified by manual inspection of the protein families and comparing this output with entries in the Clusters of Orthologous Groups of proteins (COGs) database¹³. The manually curated COGs database is based on genome-specific best hits [Tatusov et al., 2003]. Overall the output generated by MCL was comparable to COGs (data not shown). The protein families of interest were then analysed further, as described in chapter 4.

3.2.4 Functional mapping of GO terms to the SubtiList classification hierarchy

Following the classification of the functionally related secreted protein families, the properties of these families were then investigated. Of particular importance was the desire to summarise families in terms of the function of their member proteins. The SubtiList¹⁴ classification codes (shown in appendix B) provided a commonly used basis on which to define the function of *B. subtilis* proteins that is well understood by the *Bacillus* community and devised by the creators of the SubtiList database. SubtiList is a relational database for the genome sequence of *B. subtilis* (strain 168)

¹³COGs Website: <http://www.ncbi.nlm.nih.gov/COG/>

¹⁴SubtiList Website: <http://genolist.pasteur.fr/SubtiList/>

[Moszer et al., 2002]. The advantage of using SubtiList codes is that they are well known to biologists, have been manually curated, and contains an appropriate number of categories for the summation. However, as *B. subtilis* is the only organism whose genes have associated SubtiList classification codes, a method of inferring these codes for the genes of the other *Bacillus* species was required.

The key was the availability of Gene Ontology¹⁵ (GO) terms, assigned to individual genes, and available from the EMBL record. GO is one of the most important ontologies within bioinformatics. The directed acyclic graph (DAG) structure means that GO terms change from being very broad near the root of the DAG to more specific as the DAG is descended. GO is split into three domains:

- *Molecular function* describes activities (e.g. catalytic activity, transporter activity or binding) that occur at the molecular level performed by individual gene products (or possibly by assembled complexes of gene products).
- *Biological process* is a series of events or molecular functions, not necessarily a pathway.
- *Cellular component* describes the location of a gene product at different levels of detail, from subcellular structures to macromolecular complexes [GO Consortium, <http://www.geneontology.org/GO.doc.shtml>].

As SubtiList classification codes describe the processes involved in cellular function, the most appropriate of the GO categories for mapping to SubtiList was "biological process". The number of GO terms defining the biological processes of gene products in the 12 secretomes is large, therefore a method of summarising these GO process annotations for the protein families was therefore required.

The GO terms of the genes encoding the proteins in each protein family were examined and then summarised by classifying the terms according to the SubtiList classification codes; the GO terms were effectively mapped to the SubtiList classification codes. This was initially a manual process, but due to the growing size of the dataset, an automated approach was required.

¹⁵The Gene Ontology: <http://www.geneontology.org/>

The automated approach developed made use of a program developed by P. Lord, called go-graph. This program uses probabilities to locate a common level in the GO DAG. It is based on an *information content* measure of semantic similarity, in which the less commonly used terms are more informative [Lord et al., 2003].

In order to adapt go-graph to summarise GO process terms using SubtiList classification codes, additional plugin code was written. Firstly, a script was written to calculate the number occurrences of each of the GO process terms used to define the 12 *Bacillus* proteomes. A probability was subsequently inferred for each of these GO terms based on the number of occurrences. Using these probability scores, a smaller set of GO process terms were calculated, so as to summarise the biological processes of the *Bacillus* proteomes more effectively. Calculating a subset of GO process terms using probabilities avoids the incorporation of those terms that are very common (higher probability scores), and hence uninformative. To determine this reduced set of GO terms, a cutoff was applied to the probability scores. The most appropriate cutoff was determined based on the number of terms returned in the subset, taking into account the proportion of proteins that were uncategorised as a result i.e. those proteins annotated with GO terms having a probability greater than the cutoff probability. The threshold chosen for this purpose was a probability of 0.01. Further consideration of this subset revealed that some terms that fall outside the cutoff probability (higher probability) do not have any associated children in annotated *Bacillus* proteome. Therefore, for completeness these *Bacillus* leaf nodes were also included. This increased the number of terms included in the subset from 26 to 135. Following the calculation of this subset of GO terms, additional plugin code was written to calculate how many of members of the subset of GO terms were found in a selected group of protein families (e.g. the entire group of secreted protein families or perhaps the core families). A final script mapped these summarised GO terms, along with their associated count, to the SubtiList classification codes. Each SubtiList classification code subsequently had an associated count, defining how many proteins were annotated as being involved in that process for the specific group of families considered.

The go-graph plugin is a standalone program, separate from the classification and analysis workflows. This program was designed to simplify and automate the process of analysing the resulting data from the analysis workflow. It therefore was not incorporated into the BaSPP workflow system.

3.3 Discussion

Using a combination of *my*Grid and Taverna, a novel e-science based package (set of workflows) has been developed that performs all the tasks required to predict the final location (i.e. cytoplasm, membrane, membrane-associated, wall-associated, secreted) of each encoded protein. The workflows constructed enable secretory protein prediction over bacterial genomes using multiple prediction tools, integrates the results into a database, and then performs analysis on the families. This approach orchestrates existing bioinformatics programs in order to make predictions, which would otherwise take several days if performed manually. In particular the ease by which a workflow may be re-run as and when new genomes are sequenced is a distinct advantage, especially as the rate of complete genome sequencing continues to increase. Whilst building a workflow and the associated services is a time consuming task, the time saved in the analysis of each genome compensates for this overhead. Completion of the entire workflow took approximately 2-3 hours for each *Bacillus* proteome, which range from 4202 to 5412 proteins in size.

Another advantage of the BaSPP system is that the generated data is stored in a custom database, making the results available for subsequent analysis. The system uses this database to provide a way of sharing data and promoting collaboration through an accessible and user-friendly Web interface. This approach is different to classical workflow methods where individual users need proactive involvement in how and where results are stored.

The command line tools used in the classification phase of the workflow were initially implemented using Java Runtime and wrapped as SOAP-based RPC Web services. Due to problems encountered during the execution of the classification

workflow, BioJava was used instead, using its own runtime method. An improvement to the current implementation would be to expose the tools as Web services using Soaplab which provides a common interface to the various underlying applications.

The analysis process is the most computationally intensive section, requiring a large number of BLASTp searches. BLAST is the most commonly performed task in bioinformatics [Stevens et al., 2001], and as such there are many available services which could have been used. However, because of the computational intensive nature of large BLAST searches, pre-computed BLAST results were retrieved from the Microbase database.

Although the details of the workflow are specific to the problem of predicting secretory proteins, the architectural solutions employed highlight some general issues in e-science relating to three key issues of distribution, autonomy and heterogeneity.

Firstly, most of the services in the classification workflow are, in fact, provided originally by external parties, autonomous from the workflow authors. As has been mentioned previously [Stevens et al., 2004b], the reliability of the services deployed by many providers is not high. The problem is particularly serious in our case as the workflows are largely linear; the failure of a single service will cause the entire workflow to fail. This problem was solved by the simple expediency of hosting the services locally, although this is not ideal due to the time and effort required for their local implementation. Having developed local services, it would be appropriate to republish them for reuse as a service to the community. There is, however, the secondary problem of licensing agreements. The programs SignalP, LipoP, TMHMM are all subject to license agreements. In most cases, the services are not allowed to be exposed for use by non-licensed individuals.

There were few problems introduced by distribution. Local installation of programs meant there were few network effects due to data transfer. The most significant recurrent difficulty came from the relatively large datasets that were being dealt with. This was one of the motivations for the linear shape of the classification workflow. Two more principled approaches to this problem can be seen. Firstly, improved data transport facilities, enabling transfer without SOAP packaging, as well as direct

transfer between third parties, would reduce many difficulties. The new Taverna2 architecture should enable these improved data transport functionalities [Oinn and Pocock pers. comm.]. Secondly, the ability to migrate workflows and services closer to the data would provide a significant advantage; as the data are large, and the executables small, moving the program to the data would make sense.

Data heterogeneity provided fewer problems than expected for a bioinformatics workflow. Relatively few data formats were employed in the workflows. Most often simple FASTA formatted protein lists are passed between the services in the workflows. Data heterogeneity was dealt with through the use of a custom database to warehouse data and parsing code to populate this database. In the second workflow, the use of the Microbase data warehouse avoided many heterogeneity related problems, simplifying the process of data analysis. The comparison of the bacterial proteins had already been done by Microbase, thereby reducing the computing time required in protein family construction.

The architecture of the BaSPP system differs somewhat to most of the other ^{my}Grid workflow-based systems described previously. Whilst the original requirements appeared to favour a highly distributed approach, technological and licensing constraints have led to a hybrid approach: a combination of services and workflows, combined with databases for both results storage and pre-caching of analyses. The combination of all of these technologies results in a fairly complex architecture but provides a system that is fast and reliable enough for practical use. In addition, the shape of the workflows constructed are linear, as opposed to fanned/tree-like. This perhaps reflects the nature of the analysis being performed, in which the results of one service effects the input to proceeding services. Compare this to Graves' Disease in which public databases such as EMBL, GO, HGVBASE and MEDLINE are queried in tandem in order to extract information about gene structure and function, chromosome location, the presence of single nucleotide polymorphisms (SNPs), expression control features and association with other genetic diseases. In addition, w

To facilitate the process of the functional annotation of protein families, an automatic method of summarising the functional distribution using an appropriate number

of categories was required. GO terms provided the basis for this, as GO terms were annotated in the EMBL record for individual genes. However, as GO is a DAG, in which GO terms change from being very broad near the root of the DAG to more specific as the DAG is descended, the obvious way of doing this was to set a cutoff at a specific level in the DAG. The problem with this approach is that gene products are annotated at different levels in the DAG. This problem can be illustrated by considering a protein family in which all 36 proteins have a related biological process GO term. 32 out of 36 proteins are annotated as *transport*, whereas the remaining 4 are annotated as *amino acid transport*. As highlighted in the screenshot taken from AmiGO¹⁶ (figure 3.15), both terms occur at different levels in the GO hierarchy; the term GO:0006810 *transport* is less specific than the term GO:0006865 *amino acid transport*.

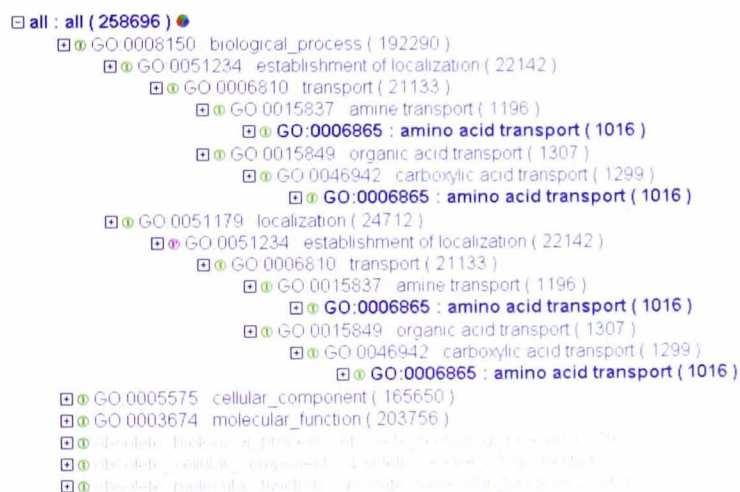


Figure 3.15: An example of the GO hierarchy, as displayed by AmiGO.

A possible solution to the annotation problem was to use a subset of GO terms (known as GO slims) providing a broad overview, without the fine-grained detail [GO Consortium, <http://www.geneontology.org/GO.slims.shtml>]. However, it became ap-

¹⁶AmiGO Website: <http://www.genedb.org/amigo/perl/go.cgi>

parent that the GO slim categories were too vague. While the construction of a new set of categories specific to secretion would combat this issue, the SubtiList classification codes turned out to provide an appropriate alternative.

The mapping of GO terms to the SubtiList classification codes was beneficial in that these annotations are known by biologists. However, these terms do not entirely capture the functional annotations of the secreted proteins. Ideally, a new set of terms, specific to the *Bacillus* species, and more specifically to the secretome, could be defined and used to provide a more definitive summary of the functions.

Despite the optimisation of BaSPP for the analysis of members of the genus *Bacillus*, the package can easily be adapted for application to other Gram-positive bacteria and even Gram-negative bacteria. Using the Web interface the resulting data can also be viewed and curated. In fact all publicly available complete genomes for the Gram-positive bacteria *Staphylococcus aureus* were categorised for use in further studies.

Finally, whilst BaSPP provides an interesting and new example of an e-science-based application, the results generated are of course of great interest to microbiologists. The results of categorisation show a high degree of correlation to comparable bioinformatics analysis and validated experimental data. This work therefore provides data to prime further biologically- and computationally-oriented investigations. Analysis of the data therefore became the focus of this study, and one to which the remaining chapters are dedicated.

Chapter 4

Properties of the predicted secretomes of members of the genus *Bacillus*

4.1 Introduction

In this study, the BaSPP system was specifically applied to the *Bacillus* genus, in which, as well as *B. subtilis* (strain 168), an additional 11 *Bacillus* strains, listed in section 1.6.2, for which complete genomic sequences were publicly available from Genome Reviews, were analysed. This builds on previous studies that have largely focussed on the model Gram-positive bacteria, *B. subtilis* (strain 168) [Antelmann et al., 2001, Tjalsma et al., 2004, 2000]. Following the categorisation and analysis of the secretomes by the BaSPP workflow system, the relationships between the secretome components were investigated. The different bacterial secretomes were compared to reveal potentially interesting differences between species. The numbers of proteins categorised as being secreted by the classification workflow were initially assessed in terms of their distribution across the various *Bacillus* species. The functionally related families identified by clustering these secreted proteins were then used to further investigate relationships between the different species based on their secretome composition. Those protein families containing members from all 12 genomes, and those specific to the non-pathogens and known-pathogens, were then investigated to identify functional trends.

4.2 Results

4.2.1 Overview of the predicted secretome composition of 12 *Bacillus* genomes

The initial findings from the classification workflow (described in section 3.2.2) are highlighted in figure 4.1. It can be seen that the majority of the secreted proteins are classified as SpI proteins, followed by SpII, and finally proteins covalently attached to the cell wall. Of the organisms used in this study, *B. anthracis* (Sterne), *B. anthracis* (Ames ancestor), *B. anthracis* (Ames), *B. cereus* (E33L) and *B. thuringiensis* *konkukian* (strain 97-27) are considered known pathogens. The remaining species *B. clausii* (KSM-K16), *B. halodurans* (C-125), *B. licheniformis* (ATCC 14580, sub_strain Novozymes), *B. licheniformis* (ATCC 14580, sub_strain Goettingen) and *B. subtilis* (strain 168) form the non-pathogens. The pathogenicity of *B. cereus* (ATCC 10987) and *B. cereus* (ATCC 14579) is unclear and were therefore not placed in either the known pathogen group or the non-pathogen group. In regards to the secretomes of these two sets of species, the proportion of proteins that are secreted is higher in pathogens than in non-pathogens (table 4.1 and figure 4.2.a), but is proportional to the genome size (figure 4.2.b). The predicted secretomes varied in size from 358 proteins in *B. clausii* (KSM-K16) (9% of the total proteome) up to 508 proteins in *B. cereus* (E33L) (10% of the total proteome).

Interestingly, of the seven proteins predicted to be covalently attached to the cell wall, one is found in each of the five known pathogens: *B. anthracis* (Sterne), *B. anthracis* (Ames ancestor), *B. anthracis* (Ames), *B. cereus* (E33L) and *B. thuringiensis* *konkukian* (strain 97-27), with the remaining two proteins belonging to *B. cereus* (ATCC 10987). These cell wall proteins may therefore play a role in virulence, and therefore further investigation is required.

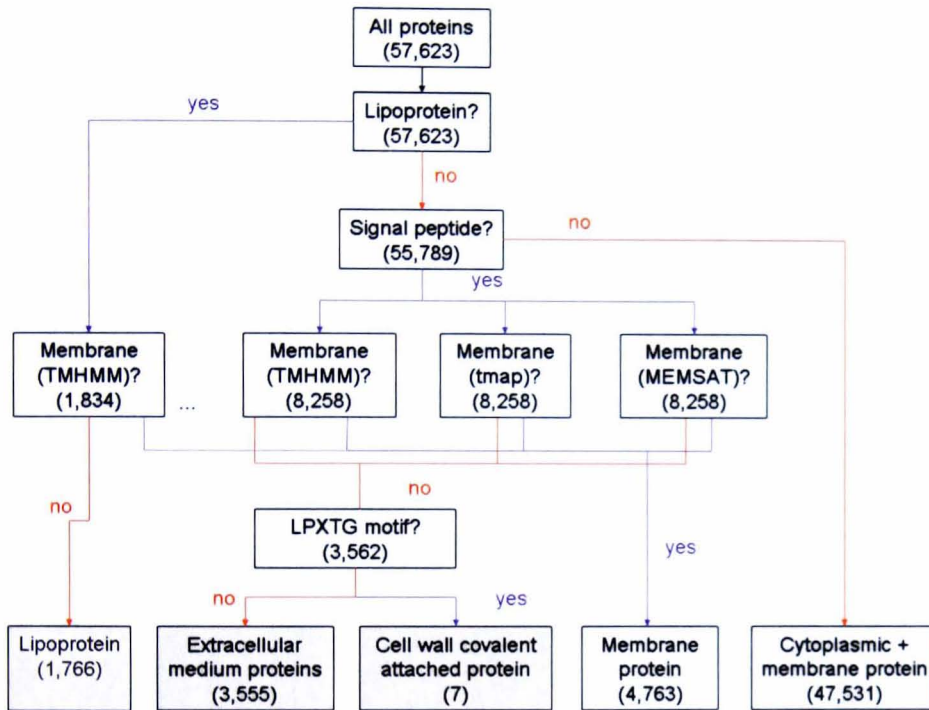


Figure 4.1: Basic representation of the functionality of the classification workflow. Shaded boxes indicate the set of secreted proteins i.e. those extracellular proteins that are secreted to the culture medium, as well as the membrane-attached lipoproteins and the cell wall covalently attached proteins. The number in brackets represents the total number of proteins to be classified at each level, across the 12 *Bacillus* proteomes.

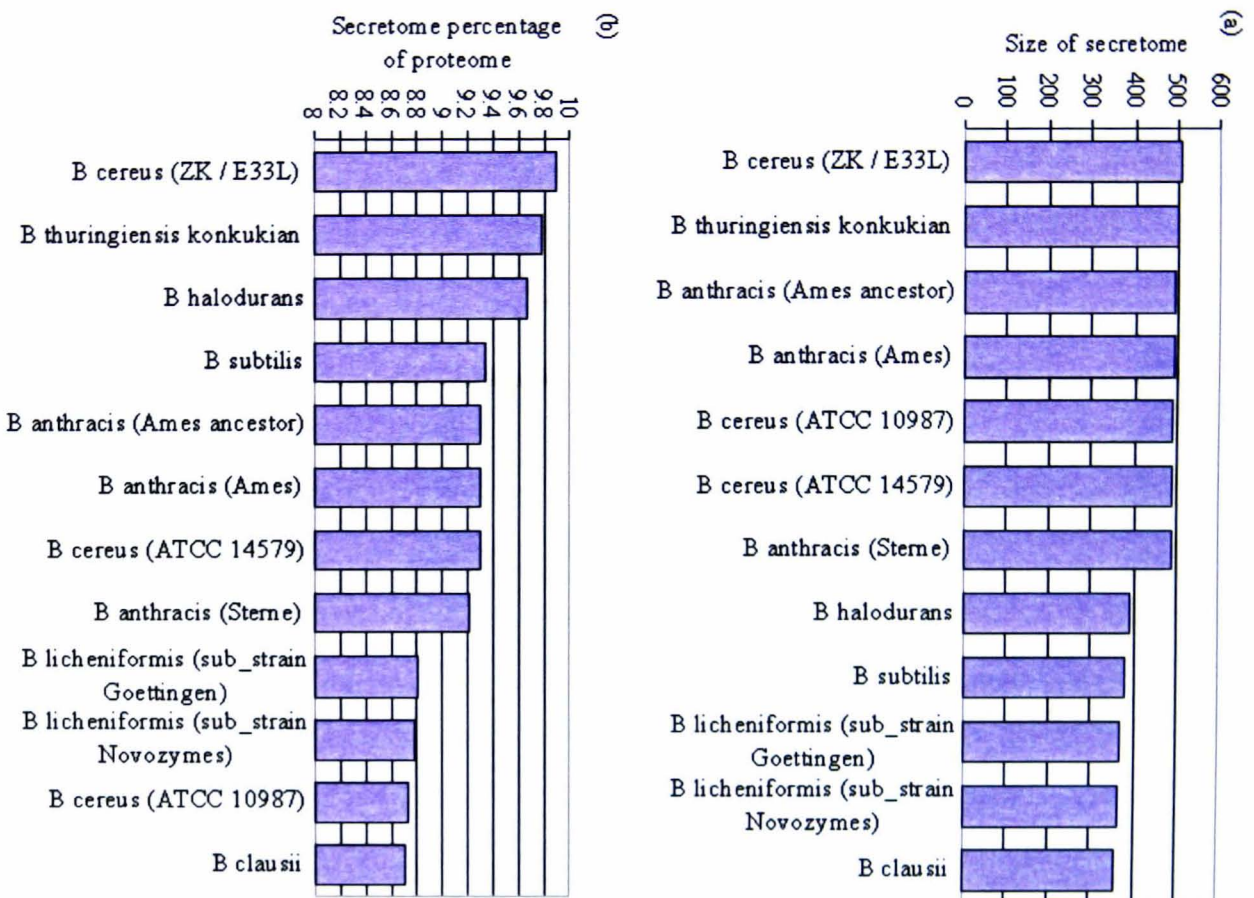


Figure 4.2: Comparison of the size of the secretomes based on (a) the total number of proteins contributing to each secretome and (b) the percentage of the each proteome classified in the secretome.

	Genome size (Mb)	Proteome size	Secretome size
<i>B. anthracis</i> (Sterne)	5.229	5287	487(9.21%)
<i>B. anthracis</i> (Ames ancestor)	5.227	5309	494(9.30%)
<i>B. anthracis</i> (Ames)	5.227	5311	494(9.30%)
<i>B. cereus</i> (E33L)	5.301	5134	508 (9.89%)
<i>B. cereus</i> (ATCC 10987)	5.224	5603	489(8.73%)
<i>B. cereus</i> (ATCC 14579)	5.412	5234	487(9.30%)
<i>B. clausii</i> (KSM-K16)	4.304	4108	358(8.71%)
<i>B. halodurans</i> (C-125)	4.202	4066	393(9.67%)
<i>B. licheniformis</i> (ATCC 14580, sub_strain Novozymes)	4.222	4152	365(8.79%)
<i>B. licheniformis</i> (ATCC 14580, sub_strain Goettingen)	4.223	4196	370(8.82%)
<i>B. subtilis</i> (strain 168)	4.215	4106	383(9.33%)
<i>B. thuringiensis konkukian</i> (strain 97-27)	5.238	5117	501(9.79%)

Table 4.1: Genome size (number of base pairs), proteome and secretome sizes (number of proteins) of each of the bacterial strains in the study.

4.2.2 Taxonomic similarity between the *Bacillus* species based on predicted secretome composition

To determine whether the secretomes of the pathogenic organisms were more closely related to each other in terms of their activity than to those of the non-pathogens, the phylogeny of the *Bacillus* strains was investigated in the context of the relationships between their secretomes. The analysis workflow was used for this purpose (described in section 3.2.3) in which the predicted secreted proteins of each strain were classified into functionally related families based on their sequence similarity, as defined by BLASTp. This resulted in the 5329 secreted proteins of the 12 secretomes being arranged into 673 families of two or more members. 618 of these proteins showed no significant similarity to other proteins and hence did not fall into any families.

As part of the output generated by this workflow, a dendrogram was produced (illustrated in figure 4.3) in which the relation between the different *Bacillus* strains is shown based on the contribution of the predicted secreted protein families. The dendrogram essentially highlights the level of similarity between the secretomes of

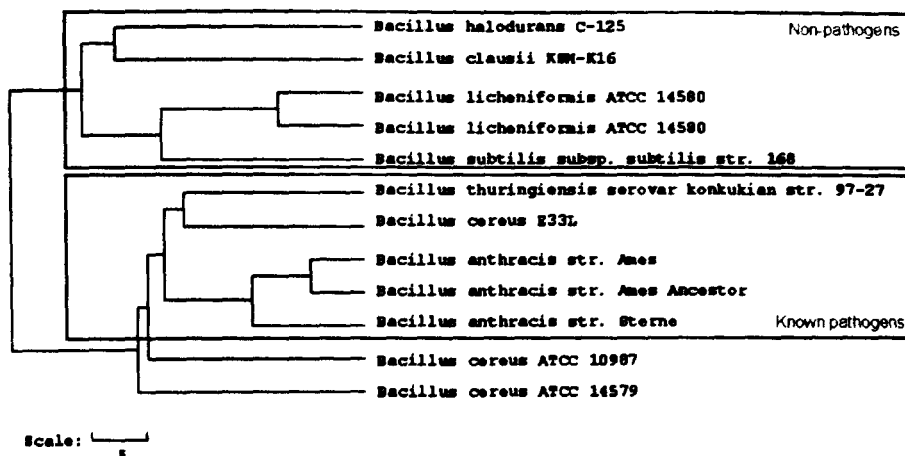


Figure 4.3: Dendrogram representing the relationship of the *Bacillus* species in terms of their secretome.

the various strains.

Within the *B. cereus* group subcluster (*B. anthracis*, *B. cereus* and *B. thuringiensis*), two sub-clusters were formed by the well-established pathogens (*B. cereus* (E33L), *B. thuringiensis konkukian* (strain 97-27), *B. anthracis* (Ames), *B. anthracis* (Ames ancestor), *B. anthracis* (Sterne)), while the two members of questionable pathogenesis (*B. cereus* (ATCC 10987), *B. cereus* (ATCC 14580)) formed a separate protein family. The environmental strains (*B. subtilis* (strain 168), *B. licheniformis* (ATCC 14580), *B. clausii* (strain KSM-K16), *B. halodurans* (C-125)) formed a separate protein family from that of the *B. cereus* group organisms.

The contribution of these organisms to the various protein families was then investigated. This is illustrated through the use of a heatmap (in a similar fashion to the visualisation of microarray data) (figure 4.4). Heatmaps are useful for visualising large and complex datasets in a way that makes it easy to compare and identify features at a glance using colour. Like the dendrogram, the heatmap was also generated using the analysis workflow. From the heatmap, trends in the distribution of the different species among the protein families can be identified. For example it can be seen that *B. halodurans* and *B. clausii* have a high proportion of paralogues specific

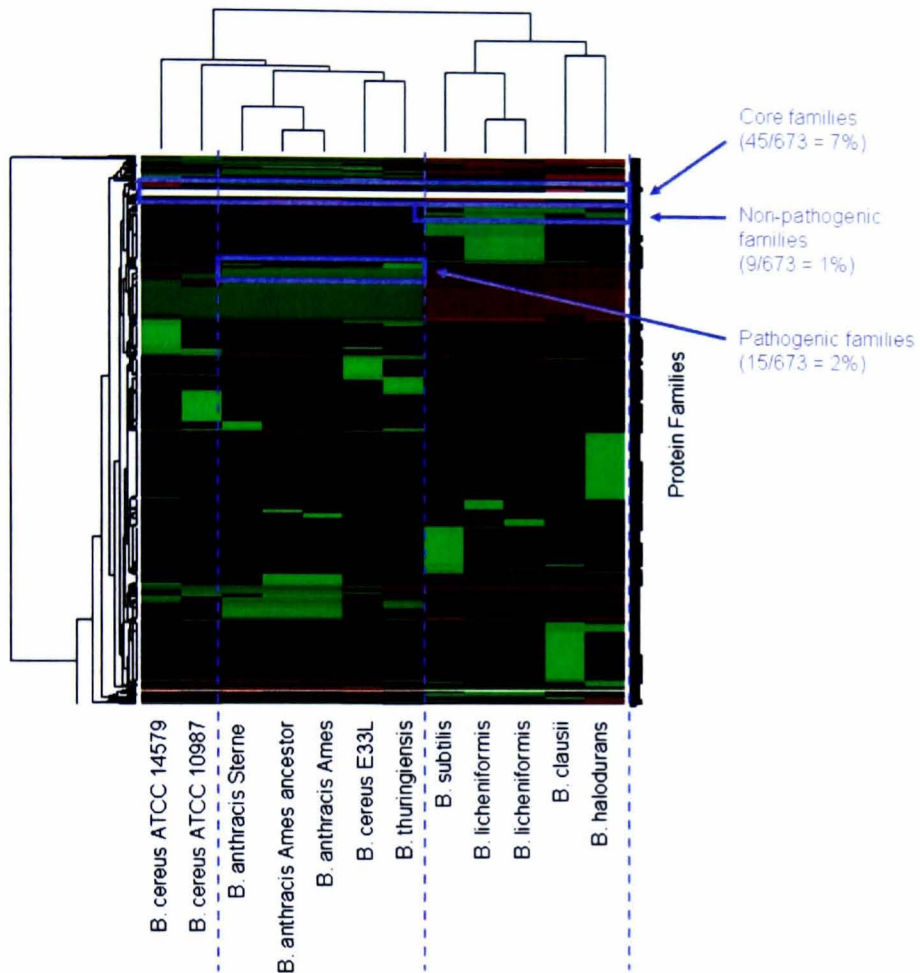


Figure 4.4: Heatmap representing the relationship of the *Bacillus* species in terms of their contribution to the protein families. The number of paralogues are taken into account for each bacteria i.e. the number of proteins from a bacteria contributing to a family. A colour gradient using three colours was used to show intensity: black shows no change in paralogue count, increasingly intense green indicates a more pronounced paralogue count in a protein family for a particular species in relation to the others, increasingly intense red indicates less paralogue count (or zero proteins) in a protein family for a particular species, white indicates the core protein families containing a member from all the species.

only to each of these species (illustrated by the green segments in the columns of these species). These species-specific families contribute to the fact that *B. halodurans* and *B. clausii* are less closely related to the other non-pathogenic species. In this study however, particular attention was paid to those families defined as "core" protein families, since they combined at least one homologue from each strain under study, as well as families specific to the pathogens and non-pathogens. Analysis of the functions of these families was subsequently undertaken.

4.2.3 Functional analysis of the predicted secretomes of the *Bacillus* species

The secretory protein families defined by BaSPP provide an opportunity to gain an overview of how the function of secreted proteins differ across the *Bacillus* species studied. The "core" families are of importance since their function is conserved and is therefore required for survival in a variety of different ecological niches. 7% of protein families were found to be core and the functions of these were investigated. Protein families that are specific to known pathogens (2%) are also of interest since they may indicate proteins whose function is related to virulence and pathogenicity. Another interesting group was the 12% of secreted proteins not clustered into any family i.e. these proteins show no relationship to any other secreted protein and hence are unique to specific strains (table 4.2).

Number of proteins analysed	57623
Number of proteins characterised as secreted	5328
Number of secreted protein families after clustering	673
<i>of which:</i>	
core secreted protein families	45
non-pathogen specific protein families	9
known-pathogen specific protein families	15
Number of secreted proteins not clustered into a family	618

Table 4.2: Summary of the number of proteins and families categorised as secreted.

The functional distribution among the protein families was investigated using the automated method described in section 3.2.4. This approach summarises the function

of the families by mapping the GO process terms associated with individual genes to the SubtiList classification codes. The subsequent functional distribution of the 12 secretomes is illustrated in figure 4.5. Inspection of the distribution across all 12 secretomes reveals a large proportion of secreted proteins have functions relating to transport and nutrient uptake. However, the majority of protein families have unknown function. This demonstrates that the secretomics of the *Bacillus* species is still not fully understood. There is therefore a great need for further studies in this area.

4.2.3.1 Core families

Figure 4.6 shows a summary of the different functional classifications of the core secreted protein families (i.e. families containing one or more proteins from each *Bacillus* species). Interestingly, a large number of core families had not been experimentally characterised and remain of unknown function. More predictably, many core proteins were grouped into families concerned with cell wall related functions, transporter proteins and proteins responsible for membrane biogenesis.

In particular, a core family of great interest is identified as being PrsA-like. PrsA is a major extracytoplasmic folding factor involved in the post-translocation folding of proteins into their native conformation after signal peptide cleavage. It has been shown to effect the function of some essential translocated proteins involved in cell wall synthesis and functions relating to the cell membrane [Tjalsma et al., 2000, Wahlström et al., 2003].

4.2.3.2 Families specific to the non-pathogen

Of the nine families unique to non-pathogens, four encoded proteins concerned with the degradation and transport of plant polysaccharides, one was concerned with the structure of flagellae, and the remaining four were functionally unclassified (table 4.3).

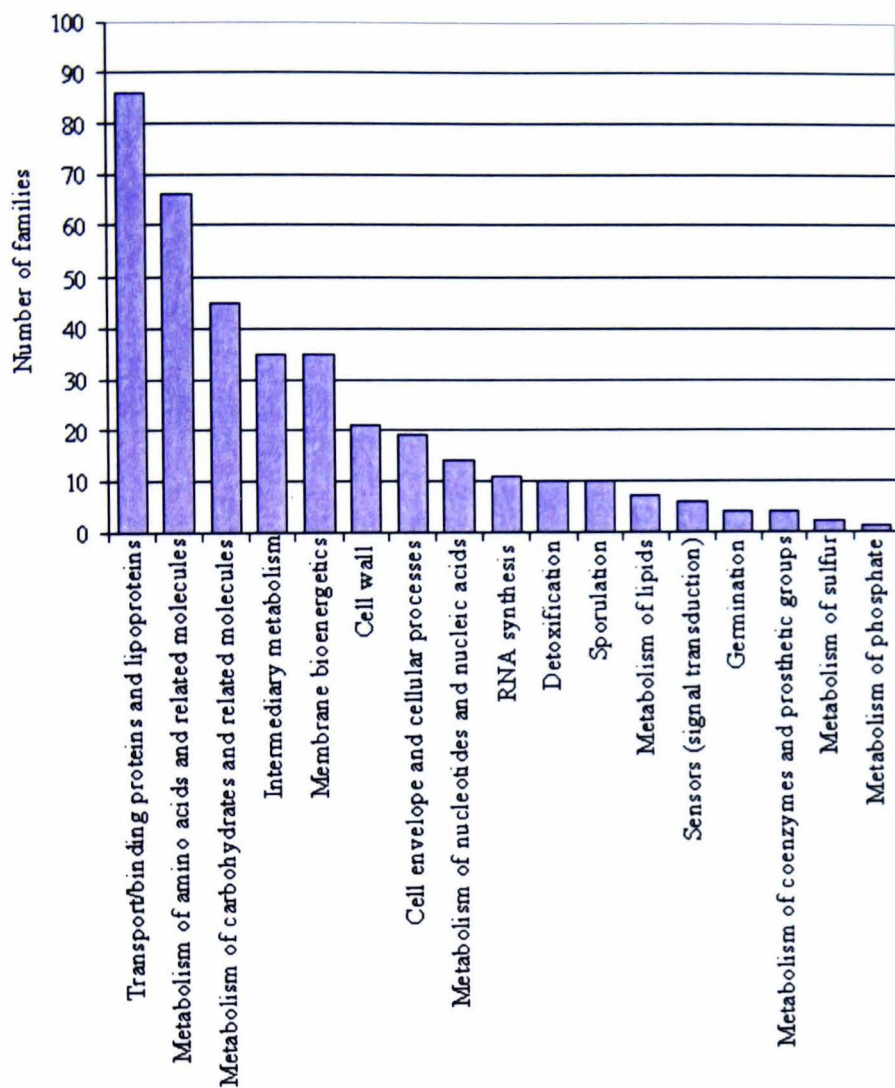


Figure 4.5: Functional classification of all the secreted protein families. The graph shows the number of protein families per SubtiList category. (Protein families classified as miscellaneous (unknown) have been excluded; this corresponds to 297 families. The GO terms mapping to miscellaneous are: GO:0008150 biological process, GO:0009058 biosynthesis, GO:0044249 cellular biosynthesis, GO:0007582 physiological process, GO:0050875 cellular physiological process, GO:0051244 regulation of cellular physiological process, GO:0050791 regulation of physiological process, GO:0009987 cellular process, GO:0050794 regulation of cellular process, GO:0050789 regulation of biological process, GO:0009605 response to external stimulus, GO:0009628 response to abiotic stimulus, GO:0042221 response to chemical stimulus.)

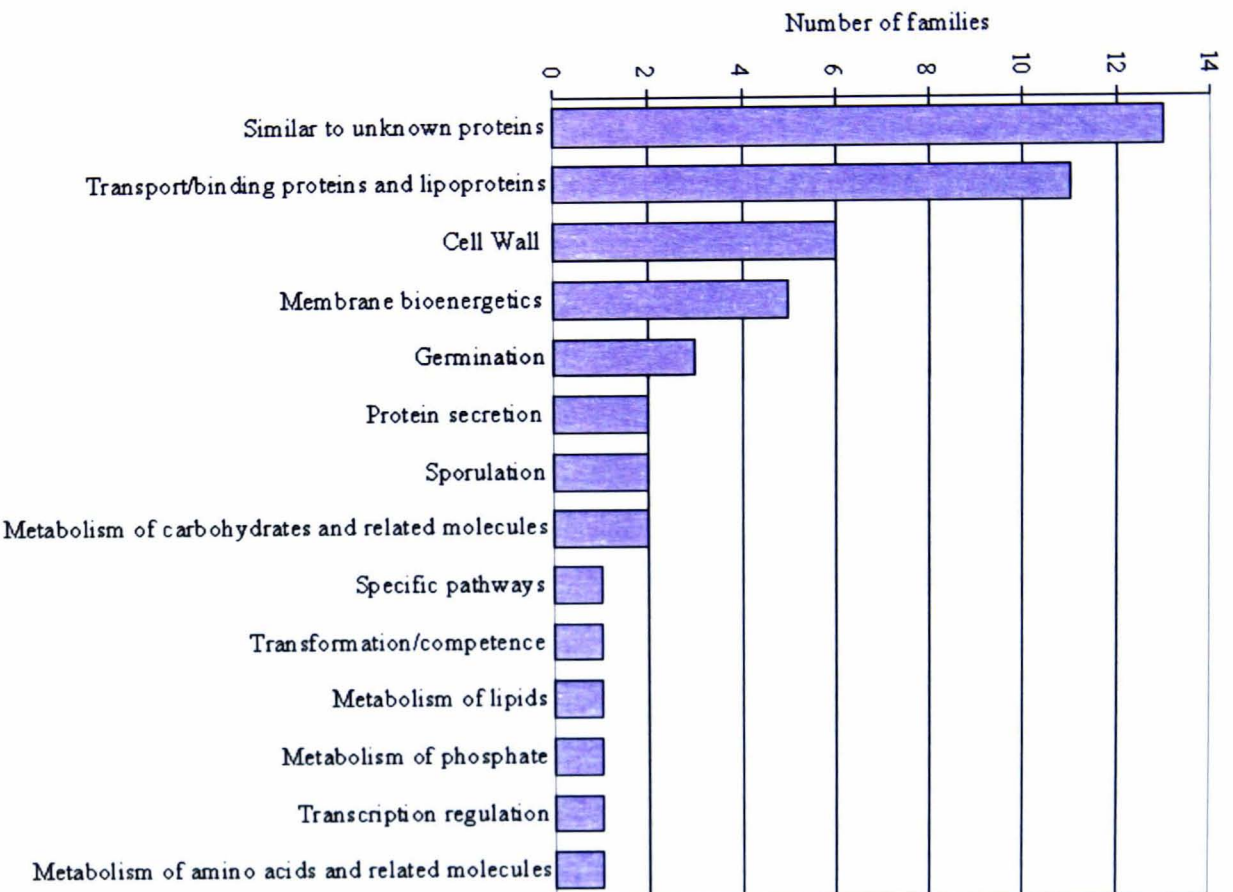


Figure 4.6: Functional classification of the core secreted protein families. The graph shows the number of core secreted proteins per Subtilist category.

Function	Num. of families annotated with function	Representative gene of the family	<i>B. clausii</i> (KSM-K16)	<i>B. halodurans</i> (C-125)	<i>B. licheniformis</i> (ATCC 14580, sub_strain Novozymes)	<i>B. licheniformis</i> (ATCC 14580, sub_strain Goettingen)	<i>B. subtilis</i> (strain 168)
Transport binding proteins and lipoproteins	3	-	ABC3357 ABC0439 ABC4079 ABC0608 ABC0301 ABC4046	BH2673 BH2750 BH3390 BH0701	BL00277	BLi02120	BSU04440
		<i>araN</i>	ABC3471 ABC3774 ABC0385 ABC3301 ABC3215	BH0905 BH1117	BL00348	BLi03024	BSU28750
		<i>lplA</i>	ABC0724 ABC3117	BH1064 BH1913	BL03778	BLi01384	BSU07100
Metabolism of amino acids and related molecules	1	<i>sppA</i>	ABC2747	BH3198	BL00425	BLi03092	BSU29530
Mobility and chemotaxis	1	<i>fliL</i>	ABC2259	BH2447	BL01263	BLi01851	BSU16300
Unknown	4	<i>pelB</i>	ABC0063	BH0698 BH3819	BL00947 BL03760 BL00361	BLi03053 BLi01404 BLi04129	BSU07560 BSU18650
		<i>yqfA</i>	ABC1672	BH1357	BL01411	BLi02729	BSU25380
		<i>ypmS</i>	ABC1528	BH1181	BL03306	BLi02310	BSU21730
		<i>ylbC</i>	ABC2386	BH2604	BL02986	BLi01713	BSU14960

Table 4.3: Functional breakdown of the clusters that only contain members from the non-pathogens. The proteins within the families are identified by their locus tags.

Function	Count
Unknown	13
Membrane bioenergetics	1
Metabolism of amino acids and related molecules	1

Table 4.4: Functional breakdown of the protein families that only contain representatives from the known pathogens.

4.2.3.3 Families specific to the known pathogen

Perhaps of most significance are the functions of the 15 protein families that were identified as being unique to the secretomes of the potentially pathogenic bacteria (*B. cereus* (E33L), *B. thuringiensis konkukian* (strain 97-27), *B. anthracis* (strain Sterne), *B. anthracis* (Ames ancestor), *B. anthracis* (Ames)) via computational means, are largely of unknown function and remain to be characterised (table 4.4).

The two families with a classified function include the membrane bioenergetics family (consisting of BA3674, BAS3405, GBAA3674, BT9727_3367, BCE33L3317) and the family classified as metabolising amino acids and related molecules (consisting of BA5606, BAS5208, GBAA5606, BT9727_5040, BCE33L5056).

Manual inspection of the 15 pathogenic families found five families showing significant similarity with proteins belonging to *B. cereus* (ATCC 14579) and *B. cereus* (ATCC 10987). Removal of these five families reduced the dataset to 10 protein families that are specific to the pathogens. The protein components of these pathogenic families are shown in table 4.5. The majority of these clusters contain members that show very close similarity to each other and are therefore well conserved (table 4.6).

Investigation into the function of the pathogen-specific families (through the application of ClustalW, BLASTp and Interproscan) indicated seven of these families have no functional motifs as defined by Interpro, and in some cases are conserved specifically within the known *Bacillus* pathogens, showing no significant similarity to any other organisms, as indicated by BLASTp.

Protein family F is an example of a protein family about which very little is known. Contained within the members of this family are three paralogues specific to *B. cereus* (E33L). The ClustalW sequence alignments for the proteins and the degree

protein family	<i>B. anthracis</i> (Ames ancestor)	<i>B. anthracis</i> (Ames)	<i>B. anthracis</i> (Sterne)	<i>B. cereus</i> (E33L)	<i>B. thuringiensis</i> <i>konkukian</i> (strain 97-27)
A	GBAA2803	BA2803	BAS2613	BCE33L2530	BT9727_2561
E	GBAA1900	BA1900	BAS1762	BCE33L1711	BT9727_1739
F	GBAA2029	BA2029	BAS1885	BCE33L1837 BCE33L2963 BCE33L4836	BT9727_3017
G	GBAA2527	BA2527	BAS2350	BCE33L0941 BCE33L2267	BT9727_0952 BT9727_2309
I	GBAA1601	BA1601	BAS1485	BCE33L1457	BT9727_1458
J	GBAA4462	BA4462	BAS4141	BCE33L3991	BT9727_3981
K	GBAA2986	BA2986	BAS2775	BCE33L2705	BT9727_2724
L	GBAA2500	BA2500	BAS2320	BCE33L2241	BT9727_2285
M	GBAA3523	BA3523	BAS3267	BCE33L3182	BT9727_3240
N	GBAA3443	BA3443	BAS3190	BCE33L3092	BT9727_3171

Table 4.5: The proteins (identified by their locus tags) belonging to the clusters that only contain members from the five known pathogens.

protein family	Percentage similarity	
	Range	Mean
A	95 - 100	97.4
E	98 - 100	98.8
F	34 - 100	59.7
G	35 - 100	69.2
I	95 - 100	97.6
J	96 - 100	97.8
K	96 - 100	97.4
L	98 - 100	99.2
M	97 - 100	98.8
N	94 - 100	96.7

Table 4.6: Measure of the percentage similarity between proteins within the known pathogen specific clusters.

of similarity between the various sequences is illustrated in figure 4.7. The sequences of the proteins in the *B. anthracis* strains were identical, with their homologues in the other strains showing varying degrees of similarity.

The lipoproteins identified in protein family G contain two *B. cereus* (E33L) paralogues and two *B. thuringiensis konkukian* (strain 97-27) paralogues. The sequence alignment and the degree of similarity between the various sequences is shown in figure 4.8. The three proteins belonging to the three *B. anthracis* strains and BT9727_2309, a paralogue of *B. thuringiensis konkukian* (strain 97-27), were identical. The paralogue of *B. cereus* (E33L) (BCE33L2267) differs only with respect to a single conserved amino acid substitution. The additional paralogues in this protein family, BCE33L0941 (*B. cereus* (E33L)) and BT9727_0952 (*B. thuringiensis konkukian* (strain 97-27)) showed reduced levels of similarity.

Little can be determined about the extracellular proteins in protein family J. A possibility is that these proteins function like Com since they show similarity to RBTH_06021, ComG operon protein 5 of *B. thuringiensis serovar israelensis* (ATCC 35646), a *Bacillus* species not analysed in this study (figure 4.10). ComG is important in transformation and DNA binding (defined in section 2.2.4.2). The *B. thuringiensis serovar israelensis* (ATCC 35646) ComG protein has an additional N-terminal segment not annotated in the *Bacillus* pathogens analysed in this study. However, closer inspection of these ORFs reveal that they may have been misidentified, in which case they too would have an additional N-terminal segment showing similarity to ComG:

MKLYFSLEEGDLKIVKCQKGS[MAE]

depending on whether [MAE] is present at the N-terminal end already. With the addition of this N-terminal segment, the *Bacillus* proteins would no longer be secreted as, like ComG of *B. thuringiensis serovar israelensis* (strain ATCC 35646), an N-terminal signal peptide would not be predicted. However, the reverse may also be true, in that the start of ComG may have been misidentified. It may therefore be the case that there is a different transformation mechanism via which competence proteins are secreted through the Sec pathway. This remains to be investigated experimentally.

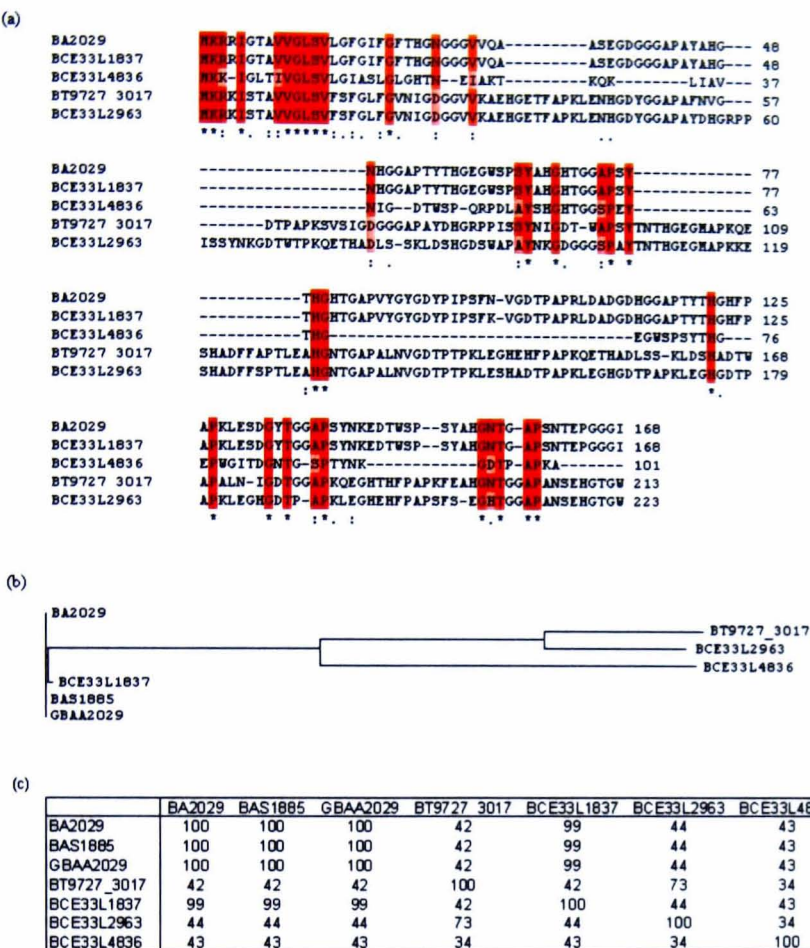


Figure 4.7: (a) ClustalW alignment for members of protein family F. As well as ClustalW symbols below the alignment ('*' means the residues in that column are identical in all sequences, ':' means that conserved substitutions have been observed, '.' means that semi-conserved substitutions are observed), the colour coding also indicates the degree of similarity (dark red refers to a match, light red means the amino acids differ but are related, white means no match). The sequences of GBAA2029 (*B. anthracis* (Ames ancestor)) and BA2029 (*B. anthracis* (Ames)) are identical; for simplicity only BA2029 is shown. Three paralogues exist for *B. cereus* (E33L), one of which shows close similarity with BT9727.3017 (*B. thuringiensis konkukian* (strain 97-27)). (b) Phylogram from ClustalW alignment highlighting the similarity between the protein sequences within protein family F. (c) The percentage identity between all sequences in protein family F.

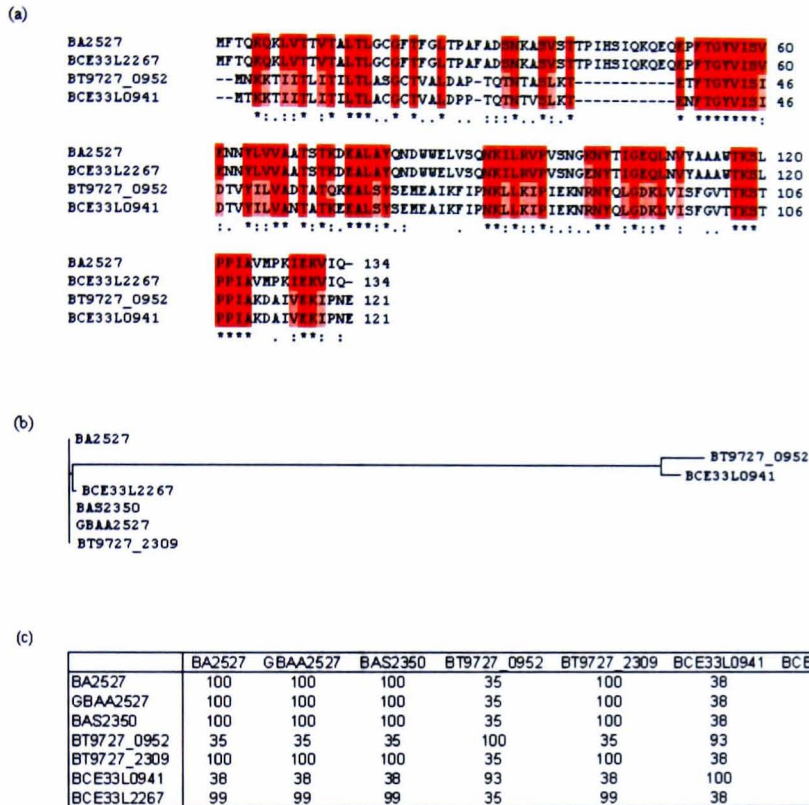


Figure 4.8: (a) ClustalW alignment for members of protein family G. Colour coding indicates the degree of similarity (dark red refers to a match, light red means the amino acids differ but are related, white means no match). The sequences GBAA2527 (*B. anthracis* (Ames ancestor)), BA2527 (*B. anthracis* (Ames)), BAS2350 (*B. anthracis* (Sterne)) and BT9727_2309 (*B. thuringiensis* *konkukian* (strain 97-27)) are identical; only BA2527 is shown for simplicity. (b) Phylogram from ClustalW alignment highlighting the similarity between the protein sequences within protein family G. (c) The percentage identity between all sequences in protein family G.

Accession	Sequence	Position
BA1601	NTKTKNLGVGVVTLSTGTLFGCAAGLFFKPNRRCBDSWDELGVWYIDBSG	60
BT9727_1458	NTKTKNLGVGVVTLSTGTLFGCAAGLFFKPNRRCBDSWDELGVWYIDBSG	60
BCE33L1457	NTKTKNLGVGVVTLSTGTLFGCAAGLFFKPNRRCBDSWDELGVWYIDBSG	60

BA1601	YFYGKTYNNKTFKNSAFKNTQBSAIFEGGIGSGNRGGGFG	103
BT9727_1458	YFYGKTYNNKTFKNSAFKNTQBSAIFEGGIGSGNRGGGFG	103
BCE33L1457	YFYGKTYNNKTFKNSAFKNTQBSAIFEGGIGSGNRGGGFG	103

A detailed manual examination of the lipoprotein clusters I (figures 4.10) and K (figure 4.11), and extracellular protein clusters L and M (figure 4.12) revealed no clues about their possible function. However, members of protein family I did show significant sequence similarity (93%) to BC1578 of *B. cereus* (ATCC 14579), but this protein is of unknown function and is not predicted to be secreted.

The protein families discussed this far (F, G, I, J, K, L, M) show no significant BLASTp results and no Interpro domains. However, analysis of the remaining three families (A, E, N) uncovered a number of interesting findings. Protein family A appears to contain genes encoding surface layer (S-layer) proteins (figure 4.13). S-

layers are crystalline assemblies of proteins that coat the outer surface of unicellular organisms ranging from the archaea, bacteria and eucarya. Their structures and functions are diverse, including acting as targets for other proteins [Sára and Sleytr, 2000].

In Gram-positive bacteria, the region responsible for the inactivation of S-layer proteins with the cell wall is termed the *surface layer homology* (SLH) domain. These domains have been found at both the N and C regions of wall-bound mature proteins. An SLH domain is usually composed of either a single or three repeating SLH motifs of approximately 50 to 60 residues. Three SLH domains were identified in the respective protein family member proteins belonging to *B. anthracis* (Ames ancestor), *B. anthracis* (Ames), *B. anthracis* (Sterne) and *B. cereus* (E33L). However, the member protein of *B. thuringiensis konkukian* (strain 97-27) was found to contain only two SLH domains, as shown in figure 4.13. It has been suggested that the SLH domains interact directly with peptidoglycan or with other cell wall-associated polymers [Navarre and Schneewind, 1999].

With the exception of *B. thuringiensis konkukian* (strain 97-27), a beta-lactamase-like region was also identified in members of protein family A. Beta-lactamases are enzymes implicated in the resistance of bacteria to beta-lactam antibiotics (e.g. penicillins). As the name suggests, beta-lactam antibiotics contain a four-membered beta-lactam ring. Beta-lactamase inactivates the antibacterial activity of beta-lactam antibiotics by cleaving the beta-lactam ring [Navarre and Schneewind, 1999, Wilke et al., 2005]. In Gram-positive bacteria, beta-lactamases are generally lipoproteins secreted by the Sec pathway and processed by SPase II. However, it has also been observed that beta-lactamases may also be processed by SPase I, but these proteins lack the ability to protect the bacteria from beta-lactam antibiotics (as observed in staphylococci). The membrane anchoring is thought to be important for resistance [Navarre and Schneewind, 1999]. These beta-lactamase-like proteins also show close similarity with the non-pathogenic *Bacillus* species, as well as *Listeria*, *Clostridium*, *Yersinia*, *Staphylococcus*, *Thermoanaerobacter* and *Streptococcus*, although these species appear to lack proteins showing the S-layer homology regions. The lack of a beta-lactamase-

like region in *B. thuringiensis konkukian* (strain 97-27) is due to a deletion of a segment of protein sequence corresponding to a length of 272 amino acids (as shown in figure 4.13).

Also of interest are the putative glycoside hydrolase proteins in protein family E, containing multiple bacterial neuraminidase repeats (BNR) or Asp-boxes (figure 4.14). Asp-boxes are found in various non-homologous proteins, such as bacterial ribonucleases, sulphite oxidases, reelin, netrins, sialidases, neuraminidases, lipoprotein receptors and glycosyl hydrolases. However, little is known about this motif, apart from the observation that they occur most frequently in secreted proteins, and are often found in proteins that interact with polysaccharides (e.g. Asp box-containing glycosyl hydrolases). O-Glycosyl hydrolases are enzymes that hydrolyse a glycosidic bond i.e. a bond between a carbohydrate and an alcohol, where the alcohol may also be a carbohydrate. These enzymes are known to regulate the activities of certain extracellular signalling proteins. However, it is unlikely that all Asp box-containing proteins will be associated with a polysaccharide-binding function. For example in barnase-like ribonucleases, the Asp box motif lies with the active site [Copley et al., 2001] suggesting a role in catalysis.

In conjunction with this finding, the proteins in protein family E also show close similarity with glycosyl hydrolase, BNR repeat-containing proteins expressed in the Gram-negative bacteria *Polaromonas sp* JS666 and *Nitrosomonas eutropha* C91. However, unlike the proteins within protein family E, the algorithms applied in the classification workflow suggest the proteins from these organisms are not lipoproteins but rather appear to be secreted.

Of particular interest are proteins within protein family N. These proteins have been identified as zinc metalloprotease enhancins. Enhancins, also known as viral enhancing factor, are pathogenicity factor encoded by *baculoviruses*. Baculoviruses are viruses that are pathogenic for invertebrates. Enhancins have been shown to enhance baculovirus infections and decrease larval survival time i.e. decreasing the survival of the pre-adult forms of an animal. It is thought that enhancins disrupt the protective peritrophic membrane, enabling the virion to attack the epithelial cells of

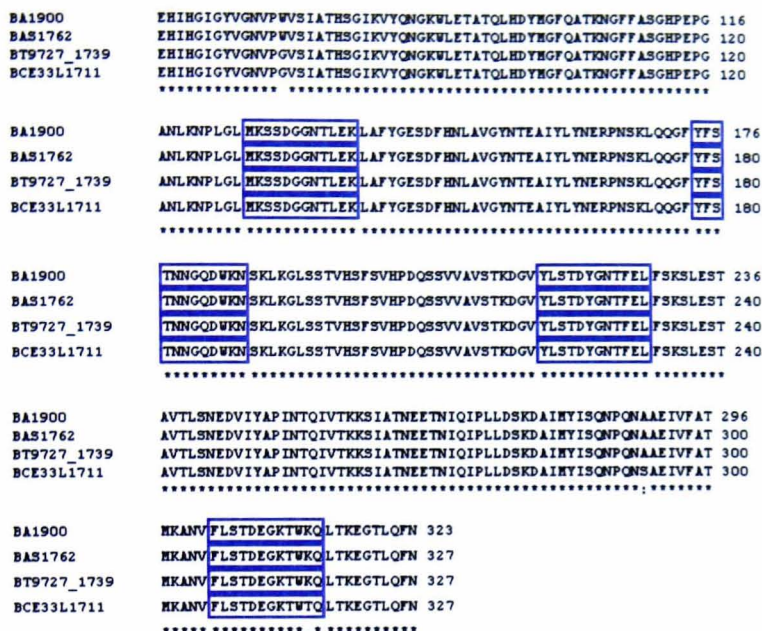


Figure 4.14: Interproscan domains mapped onto ClustalW alignment for members of protein family E. The sequences GBAA1900 (*B. anthracis* (Ames ancestor)) and BA1900 (*B. anthracis* (Ames)) are identical. Each protein within this family contains four glycoside hydrolase, BNR repeat domains (IPR002860) (blue).

the insect gut [Galloway et al., 2005, Hajaij-Ellouze et al., 2006].

Enhancin-like genes are not only limited to insect viruses, but have also been identified in bacteria, significantly in the *B. cereus* group, including *Bacillus anthracis* (Ames ancestor) (GBAA3443) [Read et al., 2003]. Interestingly, GBAA3443 is a member of protein family N. Closer inspection also reveals close similarity to Enhancin proteins in strains of Gram-negative bacteria *Yersinia pestis* and *Yersinia pseudotuberculosis*, which are also known pathogens. This is also confirmed in the literature with the identification of bacterial enhancins in *Yersinia pestis* (YPO0339) [Parkhill et al., 2001]. Like baculovirus and *B. thuringiensis*, the similar proteins of *Yersinia pestis* and *Yersinia pseudotuberculosis* are insecticidal toxins. Surprisingly, close similarity is also shown towards the BAE55343 protein of the fungus *Aspergillus oryzae*, thought to be non-pathogenic (despite some reported cases). Unlike the *Bacil-*

lus enhancins, application of SignalP and LipoP to the similar proteins of *Aspergillus oryzae* indicates they are not secreted. Less surprisingly, similarity is also shown towards BC_3384 of *B. cereus* (ATCC 14579), as also documented in [Ivanova et al., 2003]. However this protein is not predicted to be secreted, and presumably is unable to exert its toxic effect.

Enhancins have a distinct HEXXH metal binding motif. This motif is conserved in the *Bacillus* Enhancins (figure 4.16), as well as in the *Yersinia* and baculovirus Enhancins (HEIAH in bacterial Enhancins) (figure 4.17). Approximately 20 residues downstream of the first histidine within the zinc binding sequence is a conserved glutamic acid; this is seen in all Enhancins [Galloway et al., 2005]. Despite conservation of the HEXXH motif, the overall sequence similarity among the Enhancins varies between different genus (table 4.15).

	BA3443	AO090005000254	YPTB0393	LdnVgp161
BA3443	100	31	29	17
AO090005000254	31	100	46	20
YPTB0393	29	46	100	16
LdnVgp161	17	20	16	100

Figure 4.15: Measure of percentage similarity between BA3443 from *B. anthracis* (Ames), AO090005000254 from *Aspergillus oryzae*, YPTB0393 from *Yersinia pseudotuberculosis* IP 32953 and LdnVgp161 from the virus *Lymantria dispar* MNPV.

BA3443	KKRLLVIGINTYFFLIGNIHLHADER	THVKEITSLEPTWIFQAGISGKGTEDRQGLGF	60
BT9727_3171	KKRLLVIGINTYFFLIGNIHLHADER	THVKEITSLEPTWIFQAGISGKGTEDRQGLGF	60
BCE33L3092	KKRLLVIGINTYFFLIGNIHLHADER	THVKEITSLEPTWIFQAGISGKGTEDRQGLGF	60
	****	*****	
BA3443	ILQNTPLKVRQTNPHFKDLTVRLLSNDSEKESIQVGHVWIIQGGTFLVPFIIDTPTG		120
BT9727_3171	ILQNTPLKVRQTNPHFKDLTVRLLSNDSEKESIQVGHVWIIQGGTFLVPFIIDTPTG		120
BCE33L3092	ILQNTPLKVRQTNPHFKDLTVRLLSNDSEKESIQVGHVWIIQGGTFLVPFIIDTPTG		120
	*****	*****	
BA3443	EEPALLEYQVGNESATKPLPIYKQGSVSQFFSTWDQDGEYALIQGESTQLFIPKEDKE		180
BT9727_3171	EEPALLEYQVGNESATKPLPIYKQGSVSQFFSTWDQDGEYALIQGESTQLFIPKEDKE		180
BCE33L3092	EEPALLEYQVGNESATKPLPIYKQGSVSQFFSTWDQDGEYALIQGESTQLFIPKEDKE		180
	** * *****	*****	
BA3443	LVRSIKDFQSLDELIAETDIFARYDSIIGLDGSTVENKESQNRVTLKADISGAGGATYG		240
BT9727_3171	LVRSIKDFQSLDELIAETDIFARYDSIIGLDGSTVENKESQNRVTLKADISGAGGATYG		240
BCE33L3092	LVRSIKDFQSLDELIAETDIFARYDSIIGLDGSTVENKESQNRVTLKADISGAGGATYG		240
	*****	*****	
BA3443	ANVTANSTDSKRWLDKLSWGLDIA	NGYQAGFDHGGIFTGVVSNHLFQVGYQTSKYG	300
BT9727_3171	ANVTANSTDSKRWLDKLSWGLDIA	NGYQAGFDHGGIFTGVVSNHLFQVGYQTSKYG	300
BCE33L3092	ANVTANSTDSKRWLDKLSWGLDIA	NGYQAGFDHGGIFTGVVSNHLFQVGYQTSKYG	300
	*****	*****	
BA3443	KADQVGLFNFYKKKEQVERNLVYALNKENQNTDLDLRQKILLTRAKQKGADEAFARKY		360
BT9727_3171	KADQVGLFNFYKKKEQVERNLVYALNKENQNTDLDLRQKILLTRAKQKGADEAFARKY		360
BCE33L3092	KADQVGLFNFYKKKEQVERNLVYALNKENQNTDLDLRQKILLTRAKQKGADEAFARKY		360
	*****	*****	
BA3443	QGYRLASNAAFKKGDHSLPDLNNQYTSENVQVDFTPVFERUGFKLWKKQIEHNRAGKTP		420
BT9727_3171	QGYRLASNAAFKKGDHSLPDLNNQYTSENVQVDFTPVFERUGFKLWKKQIEHNRAGKTP		420
BCE33L3092	QGYRLASNAAFKKGDHSLPDLNNQYTSENVQVDFTPVFERUGFKLWKKQIEHNRAGKTP		420
	*****	*****	
BA3443	AVTSLATVFPESQLAKARALVDPDIPINSNFIVTNQQLASLGLKQHLNHLNTHNIDTL		480
BT9727_3171	AVTSLATVFPESQLAKARALVDPDIPINSNFIVTNQQLASLGLKQHLNHLNTHNIDTL		480
BCE33L3092	AVTSLATVFPESQLAKARALVDPDIPINSNFIVTNQQLASLGLKQHLNHLNTHNIDTL		480
	*****	*****	
I			
BA3443	KGKIKLKEGNTVIQKTIETADINLQDVPNGIYTVISGGKTDSETHFSTYATVKEKD		540
BT9727_3171	KGKIKLKEGNTVIQKTIETADINLQDVPNGIYTVISGGKTDSETHFSTYATVKEKD		540
BCE33L3092	KGKIKLKEGNTVIQKTIETADINLQDVPNGIYTVISGGKTDSETHFSTYATVKEKD		540
	*****	*****	
BA3443	NSLTIDVNEKVSNNLVNETIQFLGLGDDQFAELNTDLEQKRAVFTVTTKTPBSTYAGEKY		600
BT9727_3171	NSLTIDVNEKVSNNLVNETIQFLGLGDDQFAELNTDLEQKRAVFTVTTKTPBSTYAGEKY		600
BCE33L3092	NSLTIDVNEKVSNNLVNETIQFLGLGDDQFAELNTDLEQKRAVFTVTTKTPBSTYAGEKY		600
	*****	*****	
BA3443	ASIELFNEKGERITTKENEGTNTVVKDIIPLEKGYRIKITHDEVKRLTSKATINPNN		660
BT9727_3171	ASIELFNEKGERITTKENEGTNTVVKDIIPLEKGYRIKITHDEVKRLTSKATINPNN		660
BCE33L3092	ASIELFNEKGERITTKENEGTNTVVKDIIPLEKGYRIKITHDEVKRLTSKATINPNN		660
	*****	*****	
BA3443	KTNEIINTKUGLKNNTYLQNNPEENLKRIDEKEHVIISNPLKEIPBQKLEKKNQVWIAI		720
BT9727_3171	KTNEIINTKUGLKNNTYLQNNPEENLKRIDEKEHVIISNPLKEIPBQKLEKKNQVWIAI		720
BCE33L3092	KTNEIINTKUGLKNNTYLQNNPEENLKRIDEKEHVIISNPLKEIPBQKLEKKNQVWIAI		720
	*****	*****	
BA3443	NKLSEPOKITTYINKYDSLTYNE	742	
BT9727_3171	NKLSEPOKITTYINKYDSLTYNE	742	
BCE33L3092	NKLSEPOKITTYINKYDSLTYNE	742	
	*****	*****	

Figure 4.16: Interproscan domains mapped onto ClustalW alignment for members of protein family N. The sequences of BA33190 (*B. anthracis* (Sterne)), GBAA3443 (*B. anthracis* (Ames ancestor)) and BA3443 (*B. anthracis* (Ames)) are identical. Peptidase M60, viral enhancin protein motif (IPR004954) is identified (blue), including the HEIAH metal binding motif: the metal ligands (orange), the active site (red), and the conserved glutamic acid (green). All sequences are identified except BCE33L3092 (*B. cereus* (E33L)).

... continued

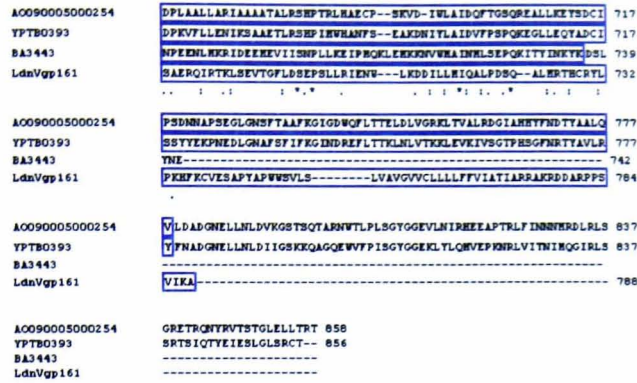


Figure 4.17: Interproscan domains mapped onto ClustalW alignment for members of protein family N. The sequences BA3443 from *B. anthracis* (Ames), A0090005000254 from *Aspergillus oryzae*, YPTB0393 from *Yersinia pseudotuberculosis* IP 32953 and LdnVgp161 from the virus *Lymantria dispar* MNPV. Peptidase M60, viral enhancin protein motif (IPR004954) is shown (blue), including the HEXXH metal binding motif: the metal ligands (orange), the active site (red), and the conserved glutamic acid (green).

4.3 Discussion

This section presents an analysis of the predicted secretomes of 12 bacilli based on their genome sequences, carried out with the BaSPP workflow based system.

Firstly, the overall composition of each individual secretome was determined. The number of proteins targeted by SpI, SpII and sortase specific mechanisms were predicted for all of the 12 isolates. Whilst the number of secreted proteins was higher in pathogenic organisms such as *B. anthracis*, the genomes of these isolates is larger, and once corrected for genome size, the number of secreted proteins remained roughly constant at around 9-10% of the total predicted proteome. These findings are in line with previously reported findings for Gram-positive bacteria.

However, it may have been anticipated that the BaSPP classification workflow would have identified more cell wall covalently attached proteins among these additional *Bacillus* species. Some may have been filtered as transmembrane proteins due to the hydrophobic domain following the LPXTG motif. However, further analysis

of the putative transmembrane proteins for LPXTG cell wall motifs did not result in additional cell wall covalently attached proteins being detected. Therefore, it is likely that the other sortase substrates were not found as their motif is slightly different to the standard LPXTG motif, documented in Boekhorst et al. [2005]. For this reason, additional sortase substrates in the *Bacillus* species have been identified in Boekhorst et al. [2005], that have not been uncovered in this study, which searched solely for the major motif LPXTG (table 4.7).

Genome	No. of putative sortase substrates predicted	
	[Boekhorst et al., 2005]	BaSPP
<i>B. anthracis</i> (Ames)	3	1
<i>B. anthracis</i> (Ames ancestor)	-	1
<i>B. anthracis</i> (Sterne)	-	1
<i>B. cereus</i> (ATCC 14579)	5	-
<i>B. cereus</i> (ATCC 10987)	6	2
<i>B. cereus</i> (ZK/ E33L)	-	1
<i>B. halodurans</i> (C-125)	6	-
<i>B. subtilis</i> (strain 168)	2	-
<i>B. thuringiensis konkukian</i> (strain 97-27)	-	1

Table 4.7: The number of putative sortase substrates in the *Bacillus* species as predicted by Boekhorst et al. [2005] and by the BaSPP system developed in this study.

Next, a method for the functional analysis of the secreted proteins was required to progress with the individual and comparative secretome analysis. The classification of proteins into families has been shown to be a valuable method of shedding light on protein function [Enright et al., 2003, Tatusov et al., 2003]. Secreted proteins were arranged into functional clusters based on amino acid sequence similarity using an MCL algorithm as part of the BaSPP analysis workflow. The clusters were subsequently systematically and manually analysed to provide an overview of the variability in the function of secreted protein families between the *Bacillus* species under study.

The phylogenetic relationships between the 12 bacilli isolates have been explored in the context of the similarity between their predicted secretomes as determined by the shared contribution of their proteins to the various families. This exercise allowed the similarity between the secretomes to be studied at the level of their

overall composition. The taxonomic relationships between the *Bacillus* species as defined by their secretomes agrees well with previously reported taxonomic trees generated by studies based on ribosomal RNA (rRNA) gene sequence analysis, usually 16S-rRNA genes. The genes coding RNA are frequently used in the area of molecular phylogeny to understand organismal relationships as these genes are highly conserved, performing central metabolic functions that are tightly intertwined in the fabric of the cell [Hale et al., 1995, Olsen and Woese, 1993].

Using rRNA gene sequence analysis, a number of bacterial species have been identified and classified, summarised in Joung and Côté [2002]. Comparison with these studies reveals that the findings presented here relate to, but do not match exactly, those documented elsewhere. In fact the Ribosomal Database Project (RDP)[Cole et al., 2005] and the study carried out in Xu and Côté [2003] show significant resemblance to the results described here. The phylogeny identified using RDP¹ is shown in figure 4.18 for comparison. The only substantial difference is the relation of *B. cereus* (ATCC 14579) to the known pathogen group containing *B. anthracis* and related organisms. RDP shows *B. cereus* (ATCC 14579) to be more closely related to the known pathogens than found by analysing the secretomes.

Recently, Tettelin and co-workers introduced the concept that a species can be described by the pan-genome that consists of a core set of proteins common to all isolates (around 80% of the total gene complement in streptococci for example) and a dispensable component that are shared across some species but not all (around 20%) [Tettelin et al., 2005]. The dispensable proteins in the pan-genome are less constrained by evolution and are therefore able to contribute to the fitness of an organism for growth in a particular environment. It is known that secreted proteins play an important part in niche adaptation and that the secretory proteins occur in young-intermediate aged functional modules overrepresented in the dispensable protein complement of an organism [Campillos et al., 2006]. The secretome of the genus *Bacillus* is particularly interesting to study in this respect. Whilst all the *Bacillus* spp. studied here share significant sequence similarity at a genomic level, the optimal

¹RDP Website: <http://rdp.cme.msu.edu/>

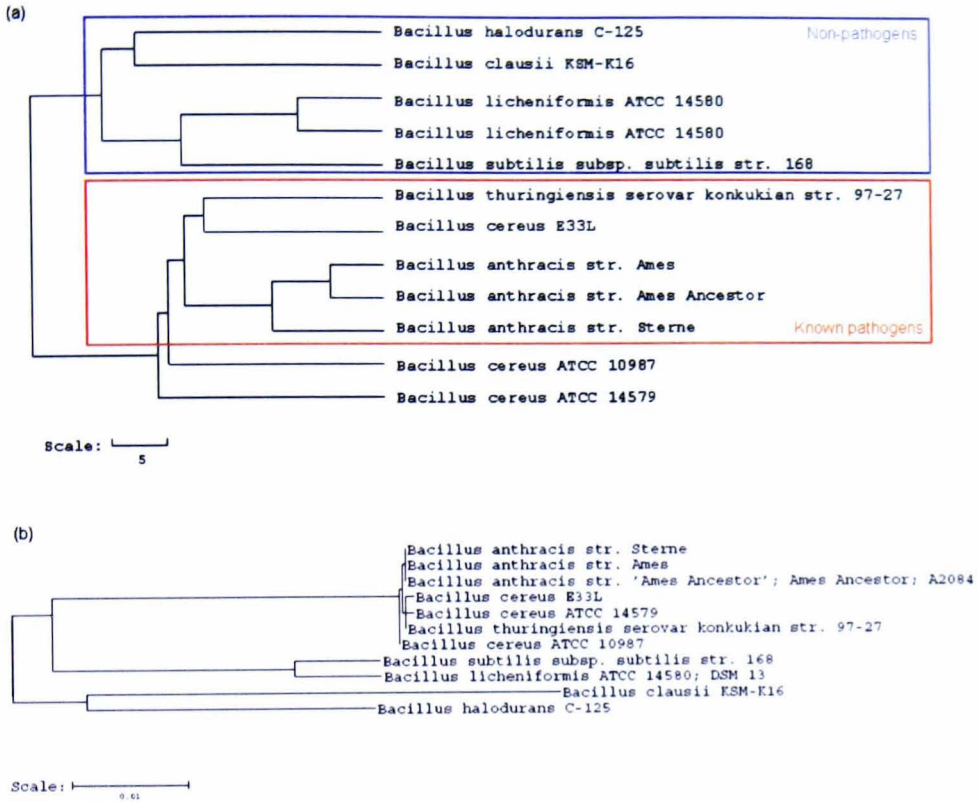


Figure 4.18: Dendrogram representing the relationship of the *Bacillus* species (a) in terms of their secretome and (b) as identified using RDP.

growth environments for these species varies tremendously from non-pathogenic soil organisms such as *B. subtilis*, through to those that are strictly pathogenic such as *B. anthracis*. The results presented here for the functional classification of the secreted protein families, with proteins of unknown function and families responsible for niche specific metabolite utilisation, support the hypothesis that secreted proteins are overrepresented in the dispensable part of the pan-genome. A more detailed set of *in-silico* experiments could be carried out to confirm these hypotheses. For instance, the pan-genome for the *Bacillus* species could be identified by clustering the

entire proteome into protein families (using the same clustering methodology). The core and dispensable components for the *Bacillus* species could subsequently be distinguished. The distribution of the secreted protein families could then be assessed by analysing the proportion of secreted protein families identified in the core and dispensable families.

Overall, the similarity between the relationships identified by analysing the secretome and rRNA shows the importance of secreted proteins, and the central role these proteins play in adapting different species to growth under different conditions. Specifically these proteins reflect the common ancestry and environmental niche in which the bacteria exist. It can therefore be speculated that the secretome of an organism would make useful biomarkers for the determination of its preferred environmental niche and properties such as virulence.

Whilst annotating these protein families, an interesting problem arose with respect to assigning functional categories to the protein families. *Bacillus* biologists are used to the functional annotations ascribed by their genome analysis portal SubtiList. In addition, when ascribing function to the families, many GO terms could be applied to a family. Simply using a broadly defined term, near the root term (as in GO slims), resulted in very vague GO annotations. Therefore the approach taken here was to map the numerous GO terms generated from a family to the existing SubtiList classification codes. This resulted in a classification system that was more human readable but less computationally amenable. A Perl-based package called go-graph was employed to extract a common set of GO terms that represented all the secreted proteins based on the probability distribution of the GO process terms in the secreted proteins dataset. This adapted version of go-graph effectively provided an automated method to determine the functional distribution across the secretomes, categorised using the SubtiList classification codes that are well-known to biologists. Automation was important due to the large number of secreted proteins that had been isolated. The application of go-graph to this task also demonstrates the advantages of using such a measure to categorise function based on the information contained within the dataset, as opposed to applying a simple cutoff level, or even using GO slims.

An examination of the occurrence of the different protein families in the different *Bacillus* species provides some interesting insights into the evolution of secreted proteins. Clearly, some protein families are shared across all of the bacilli, around 7% of the 673 protein families defined. Since this study was confined to the genus *Bacillus*, it would be interesting to determine the extent to which those families are shared across other related bacteria and even other types of organisms. The majority of the proteins classified as core are still of unknown function, suggesting there is still much to be learned about the secretory process and the role of secreted proteins. In addition, these families also generally contain predictably well conserved proteins involved in well characterised processes such as those encoding the secretory processes themselves and those involved in cell wall biogenesis and sporulation.

Of particular interest are those families that are unique to species that share phenotypic traits such as the ability to breakdown plant cell wall material and virulence. Notably, the strains reported to be able to grow in natural environments, such as *B. subtilis*, share members of families encoding enzymes responsible for the breakdown and transport of plant polysaccharides that were absent from species such as *B. anthracis* which is generally accepted not to be able to grow outside of an animal host [Jensen et al., 2003, Wipat and Harwood, 1999].

Those protein families that are associated with the pathogens have the potential to encode virulence determinants that can help with understanding the process of pathogenicity and ultimately act as the target for vaccines and antimicrobial agents. These perhaps form the most interesting targets for further laboratory investigation. Indeed, the results from this study have prompted further laboratory based experimental studies to further characterise these proteins. The definition of pathogenic strains is difficult in the absence of computational methods to describe microbial phenotypes from genotype. Many microorganisms are opportunist pathogens, not completely adapted to causing disease like a true pathogen, but able to take advantage of non-typical conditions such as immunocompromised individuals, to cause disease. This study imposed a rigid and stringent approach to the definition of pathogenicity. The virulence of the *B. cereus* strains, ATCC 14579 and ATCC 10987 has not been

proven, other than as opportunist pathogens. Both are normally soil living microorganisms known for causing food poisoning and opportunist eye infections and hence were not considered to be true pathogens.

Among the putative functional assignments for specific families were some designated as involved in cell surface activity. Perhaps of most significance was the identification of proteins that were potential bacterial enhancins in *Bacillus*, previously found in insect viruses; this is in accordance with the findings described in Read et al. [2003]. These *Bacillus* enhancins also show similarity to proteins of the known pathogens *Yersinia pestis* and *Yersinia pseudotuberculosis*. However, surprisingly this cluster of proteins was also found to be similar to a protein of the non-pathogenic *Aspergillus oryzae*. These enhancins may therefore be the result of horizontal gene transfer from insect viruses. Despite the cytotoxicity of bacterial Enhancins, the mechanism of infection differs to viral Enhancins; this may be the result of a distinct biochemical function [Galloway et al., 2005]. To confirm these findings, experiments could be devised to test for protease activity in this protein family and attempt to identify the mechanism of action. Furthermore, it would also be worth investigating the putative Enhancin activity of *Aspergillus oryzae*, not thought to be pathogenic. Investigations may uncover a pathogenic function not previously associated with *Aspergillus oryzae*, or perhaps an evolutionary relationship not considered before. It may also be interesting to identify substrate preferences of the proteins, not only for those from the *Bacillus* species, but also compared to other Enhancins.

Finally, many of the secreted proteins predicted did not cluster into families since they do not show sequence similarity to any other of the *Bacillus* secreted proteins. Presumably these proteins are strain-specific and form part of the dispensable genome discussed above.

In conclusion, the results of the BaSPP workflow have generated many novel hypotheses about the way that secreted proteins in the *Bacillus* family are used. Of course to be truly applicable, these results need to be confirmed in future laboratory based studies. In an ideal world the results of the experimental validation could be used to refine the analytical process and methodology. This system is therefore ideally

suited to an e-science approach, further demonstrating the usefulness of such an approach to bioinformatics in general.

Chapter 5

The construction of probabilistic functional integrated networks (PFINs) for members of the genus *Bacillus*

5.1 Introduction

This study aimed to identify the secretomes of the *Bacillus* species. However, as stated in section 4.2.3, a large proportion of the secreted proteins identified through BaSPP are of unknown functions. Construction of PFINs of the *Bacillus* species could provide a way to address this problem by comparing and analysing these PFINs within and across species, to provide further insight into the functions of the proteins within the *Bacillus* species in the context of their interactions. The focus was specifically on a subset of the proteomes of these bacteria, the secretomes. A comparative analysis of the *Bacillus* species, in the context of their interactomes, has not previously been undertaken. In fact, there have been relatively few reports of the application of PFINs to prokaryotes in general. The resulting framework representing all the *Bacillus* species is known as *SubtilNet*.

5.2 Methodology

The approach used to construct PFINs for the *Bacillus* species follows that documented in Lee et al. [2004] for the development of a PFIN of *S. cerevisiae* genes. This approach involves integrating a number of different datasets, each measuring different aspects of gene or protein associations, and having differing degrees of accuracy. Integration of these various datasets computationally can be achieved by realising that each experiment (genetic, biochemical, computational) adds evidence linking pairs of genes. Each dataset has different degrees of coverage (in terms of the size of the proteome) as well as different associated error rates. Functional protein-protein linkages are therefore probabilistic, providing a way of integrating the various datasets into a single interactome [Lee et al., 2004].

The generation of PFINs first requires the implementation of a unified scoring scheme to derive a consistent weighting across all experimental data sources. The scoring method relies on an additional dataset that acts as a gold standard, against which the experimental datasets are compared and weighted. Once the interactions for each experimental data source are weighted according to a common standard, the network is built by integrating all the data. An outline of the general procedures used to generate the *Bacillus* PFINs is given below.

5.2.1 Unified scoring method

Using a Bayesian approach, each experimental dataset is evaluated to determine the effectiveness of the dataset to reconstruct "known" gene pathways according to the gold standard. This procedure produces a log likelihood score, which reflects the likelihood that the pairs of genes identified by a particular experiment are functionally linked based on a gold standard. The log likelihood score reflects both the noise inherent in, and the genome coverage of the experiments, producing comparable scores across all experiments [Lee et al., 2004]. Equation 5.1 is used to calculate the log likelihood score L based on the probability that a pair of genes are either functionally linked or not functionally linked in the experimental dataset S based on the gold

standard B . $P(S|B)$ is the probability of an observed link in the dataset given their is a link in the gold standard. $P(S|\bar{B})$ is the probability of an observed link in the dataset given their is no link in the gold standard. $P(\bar{S}|B)$ is the probability of no observed link in the dataset given their is a link in the gold standard. $P(\bar{S}|\bar{B})$ is the probability of no observed link in the dataset given their is no link in the gold standard.)

$$L = \ln\left(\frac{P(S|B)/P(S|\bar{B})}{P(\bar{S}|B)/P(\bar{S}|\bar{B})}\right) \quad (5.1)$$

In order to determine all the relevant values for calculating the conditional probabilities, a matrix was constructed as shown in table 5.1.

	S	\bar{S}
B	TP	FN
\bar{B}	FP	TN

Table 5.1: Matrix required for calculation of conditional probabilities where B and \bar{B} is the number of positive and negative pairs in the benchmark respectively, and S and \bar{S} is the number of positive and negative pairs in the experimental dataset respectively.

The calculation of the log likelihood score differs depending on whether the data is weighted or unweighted. The values B and \bar{B} are constant in both cases. For unweighted data, the log likelihood score is calculated across all gene pairs within the dataset, and the same score is given to each pair associated with this dataset. However, for weighted datasets, the pairs are ranked lowest to highest weight, then bins of gene pairs are designated. The log likelihood score is calculated for each bin; using a regression plot (average weight per bin vs. log likelihood score of the pairs in that bin) the associated log likelihood score can then be obtained for each individually weighted pair. The only values to change in the log likelihood calculation for each bin are the true positives (TP) and false positives (FP). The trend in the log likelihood scores depends on the difference of these two values i.e. the lower the number of true positives, and the higher the number of false positives means the lower the log likelihood score [Lee et al., 2004].

5.2.2 Gold standard

For all of the PFINs developed, the common benchmark used to weight the various input datasets in the network was the KEGG pathway¹ dataset. KEGG is a database providing information about the functions of biological systems from genomic and molecular information. It consists of four main databases, but the pathway database is of most interest as this database provides information about protein interactions [Kanehisa et al., 2006]. Each of the 11 *Bacillus* genomes in this study are represented in this database.

All genes in the same pathway in the KEGG pathway database may be considered to be functionally related. Therefore, all pairs of genes in the same pathway for a species are linked to produce a set of positive pairs for that species. Those genes present in KEGG pathway for the same species, but not sharing a common pathway, produce the negative set of pairs for the species.

5.2.3 *B. subtilis* PFIN

A PFIN was developed for *B. subtilis* (strain 168) which served as a testbed against which the methodology described (in section 5.2) could be developed. This section describes the specific details concerning the experimental datasets and the integration process used in constructing the *B. subtilis* PFIN.

5.2.3.1 Datasets

A number of datasets were used to create the *B. subtilis* PFIN, each providing an alternative type of evidence.

- KEGG expression data

The KEGG database contains publicly available gene expression data for many different organisms, including *B. subtilis* [Kanehisa et al., 2006]. Based on this data, a matrix (gene vs. experiment) was constructed on which Pearson

¹KEGG Website: <http://www.genome.jp/kegg/>

correlation could be calculated in order to provide a weight for each predicted pair of functionally related genes.

- PubMed² cocitation abstracts

There are a number of approaches used in mining text data (e.g. Hirschman et al. [2002]). The approach employed here is similar to that described in Stapley and Benoit [2000]. Abstracts were obtained by searching PubMed for the term 'bacillus subtilis'. Using these abstracts, two genes were predicted to be functionally related if they appeared in the same abstract. Recording the number of abstracts in which a pair of genes appeared together, allowed a matrix (gene vs. gene) to be constructed. Using this matrix, pairs of genes could be weighted based on Pearson correlation.

- DBTBS operon predictions³

This dataset is described in Makita et al. [2004]. It consists of predictions as to whether two genes are functionally linked (i.e. in the same operon) based on their proximity. The prediction methodology is described in Hoon et al. [2004]. Pairs of genes within this dataset already have a weight associated with them based on Bayesian probability. By considering the weights describing the probability that a *pair of genes* occur in the same operon, predictions could be generated to determine the *cluster of genes* in the same operon. This was done by applying different thresholds to the weights defining the probability that a pair of genes belong to the same operon. Following manual inspection of the different results, it was decided a cutoff weight of 0.1 best represented the operon data. For each of these clusters, links between each gene were generated, resulting in a binary set of linkages.

- STRING fusion

STRING⁴ is a database containing integrated interaction data from known

²PubMed Website: <http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed>

³<http://bonsai.ims.u-tokyo.ac.jp/mdehoon/publications/Subtilis/operons.txt>

⁴STRING Website: <http://string.embl.de/>

and putative protein-protein interactions obtained from a number of different sources. Amongst these sources is gene fusion data. The fusion data extracted from STRING is already weighted [von Mering et al., 2007], therefore no modifications were needed in order to integrate into the *B. subtilis* network.

This use of gene fusion events to infer protein interactions relies on the assumption that if a protein is uniquely similar to two proteins in another species (i.e. their associated genes fused to produce a hybrid gene), then the component proteins most likely interact [Enright et al., 1999].

- PREDICTOME phylogenetic profiling

Predictome is a database containing putative protein-protein functional links using three computational methods based on chromosomal proximity, phylogenetic profiling and domain fusion⁵, as well as high-throughput experimental methods such as the yeast two-hybrid method. The data specific to phylogenetic profiling simply identifies putative protein-protein interactions, without an associated weight [Mellor et al., 2002].

- INTERPRO domains

Protein-protein interactions can be predicted by identifying pairs of domains enriched among known interacting proteins. Following the study described in Rhodes et al. [2005], this requires the calculation of a domain enrichment ratio (DR), which essentially quantifies the co-occurrence of particular domain pairs among interacting proteins. In order to predict domain enriched pairs, a set of known protein interactions is needed. The gold standard was used for this purpose. High DR values are strongly predictive of protein interactions, whereas smaller DR values are less strongly associated with protein interactions. The enriched domain pairs may represent physically interacting domains or indirect interactions.

To determine whether a pair of proteins interact based on the cooccurrence of

⁵Predictome Website: <http://predictome.bu.edu/static/data.html>

domains, where a common domain exists between a pair of proteins, a protein interaction was assumed based on the corresponding DR. When a pair of proteins share multiple common domains, the DR for each common domain was summed.

- BLASTp

A common approach used to identify paralogues is to take the BLAST reciprocal best hits (RBH). Applying this approach to all *B. subtilis* proteins resulted in the identification of 59 paralogue pairs. However, none of these paralogues were identified in the benchmark dataset KEGG pathway, hence the log likelihood score could not be calculated.

Therefore, paralogues were identified by analysing all reciprocal BLAST hits, as opposed to only the best hits. By using this approach, a paralogue pair was assumed for those *B. subtilis* proteins identified as showing similarity to another *B. subtilis* protein above an e-value cutoff score of $1e^{-5}$. In the case where protein A identifies protein B and protein B identifies protein A, the lowest e-value was used as the weight.

As the protein pairs identified by BLAST analysis have an associated weight (i.e. e-value), a regression plot in which the average BLASTp e-value for bins of equal size were plotted against the log likelihood score calculated for each bin. However, the scatterplot showed no trend to which a regression curve could be fitted. For this reason, the protein pairs identified via BLASTp were treated as an unweighted set. This should have limited or no effect on the final outcome, as a threshold cutoff score had already been applied to the prediction of protein pairs using BLASTp.

As a possible modification to this approach, clustering was applied to the resulting protein pairs. The clustering algorithm MCL was therefore applied, using a variety of inflation values (the parameter effecting the granularity of the clustering process), mimicking the process carried out in section 3.2.3. The

corresponding datasets resulted in worse log likelihood scores than the original dataset supplied as input. (See table D.1 in appendix D.)

The BLAST data used to infer protein pairs was simply extracted from Microbase, in which BLAST results are precomputed.

- GO Annotations (GOA)

A simple way of predicting protein interactions according to GO annotations is to use GO process terms to calculate the distance between the lowest annotated terms for two potentially interacting proteins. As GO is a DAG, the distance between two terms is considered to be the shortest path between two nodes in the graph. In addition, the lowest term may be rather broad i.e. closer to the root, and hence may not be very informative about the process in which the protein is involved. Therefore a cutoff level needs to be applied in order to eliminate unspecific terms from the predictive process.

To determine the level at which to threshold, protein pairs were generated for all proteins within *B. subtilis* whose lowest GO term was below a certain cutoff level, with the weight equal to the shortest distance between each protein pair. The cutoff levels chosen corresponded to level 3 to 11, where the root is level 1. Using the KEGG positive and negative pathway interaction datasets as gold standards, the number of TPs and FPs were calculated by comparing with the predicted protein interactions for each cutoff level. The number of the TPs versus the number of FPs was plotted to produce a ROC curve (figure C.1), from which deductions could be made regarding the optimal cutoff threshold.

It was determined from visual inspection of the graph that a minimum level of seven showed the best TP:FP ratio, therefore those protein pairs with a minimum GO annotation greater than or equal to level seven was used as input to the network. The identification of level seven as the optimal threshold level in GO agrees with the conclusions drawn in Date and Stoeckert [2006], in which they too thresholded GO level seven.

- COGs

COGs⁶ is a database containing *Clusters of Orthologous Groups of proteins* [Tatusov et al., 2003]. Proteins within the same cluster are regarded as being functionally related. Therefore this data source was used as input to the network by linking each protein in a cluster to every other gene in the same cluster, resulting in a set of protein-protein interactions for each cluster.

The log likelihood scores for each unweighted dataset (PREDICTOME phylogenetic profiling, BLASTp, COGs, DBTBS operon) and weighted dataset (STRING fusion, cocited, KEGG expression, INTERPRO domains, GOA) were calculated as described in section 5.2.1. (Data and graphs for this process are shown in appendix D.)

5.2.3.2 Integration of datasets

Following the uniform weighting of the datasets, integration of this information resulted in the formation of a *B. subtilis* PFIN. Integration of all of the experimental information concerning the linkage of a pair of genes is achieved using a naïve Bayesian approach i.e. the log likelihood score for a particular pair of genes are summed over all experiments. The various datasets were considered to be independent of each other, and therefore the degree of dependence between the various experiments was not factored into the equation.

5.2.4 Other *Bacillus* species PFINs

PFINs were developed for remaining 10 *Bacillus* species, using the same approach. As the remaining *Bacillus* genomes are less well studied, there were fewer datasets to integrate into the network. Table 5.2 summarises the datasets used.

The most significant difference from the *B. subtilis* data sources was the use of pre-computed cocitation data extracted from the STRING database. In addition, due to the specificity of DBTBS for *B. subtilis*, operon data for the other *Bacillus* species,

⁶COGs Website: <http://www.ncbi.nlm.nih.gov/COG/>

excluding *B. clausii*, were incorporated from Operon DataBase⁷. This resource contains operon data that has been extracted from the literature [Okuda et al., 2006].

Operon data was not available for *B. clausii* so gene neighbourhood information was used instead. As in STRING and Overbeek et al. [1999], gene neighbours were determined based on whether the gap between genes on the same strand were no more than 300 bases. For the clusters of genes identified, each gene was paired with every other gene within the cluster.

B. licheniformis (ATCC 14580, sub_strain Novozymes) was analysed, not *B. licheniformis* (ATCC 14580, sub_strain Goettingen) due to the lack of experimental data specific to the latter substrain.

Once the datasets had been gathered, each dataset was uniformly scored, as documented for *B. subtilis*, using the respective KEGG pathway gold standard positive and negative sets. The species-specific PFINs were constructed by integrating the appropriate data sources using the previously described naïve Bayesian approach.

5.2.5 Identification of clusters within *Bacillus* PFINs

As a first approach to analyse the PFINs, cluster analysis was performed. Clusters are highly interconnected regions within the PFINs. The clustering of proteins based on their connectivity within the PFINs can represent protein complexes and parts of functional and metabolic pathways.

Due to size of the *Bacillus* PFINs, an automatic method of clustering was needed. For this purpose, a relatively fast clustering algorithm, called MCODE [Bader and Hogue, 2003] was used. MCODE is available as a Cytoscape plugin. The MCODE parameter that most significantly influences cluster granularity is the *node score cutoff*. During the process of cluster expansion, new cluster members are only added to a cluster if their node score differs to the score of the cluster's seed node (i.e. the highest scoring node in the cluster) by less than the cutoff value. The node score cutoff is a value between 0 and 1, representing the percentage difference (e.g. a cutoff

⁷Operon DataBase (ODB) Website: <http://odb.kuicr.kyoto-u.ac.jp/>

Genomes	BLASTp	COGs	GOA	INTERPRO domains	KEGG expression	PREDICTOME phylogenetic profiling	Cocited	STRING fusion	Operons	Gene neighbourhood	KEGG pathway
<i>B. anthracis</i> (Sterne)	✓	X	✓	✓	X	X	✓	✓	✓	X	✓
<i>B. anthracis</i> (Ames ancestor)	✓	X	✓	✓	X	X	✓	✓	✓	X	✓
<i>B. anthracis</i> (Ames)	✓	X	✓	✓	X	X	✓	✓	✓	X	✓
<i>B. cereus</i> (E33L)	✓	X	✓	✓	X	X	✓	✓	✓	X	✓
<i>B. cereus</i> (ATCC 10987)	✓	X	✓	✓	X	X	✓	✓	✓	X	✓
<i>B. cereus</i> (ATCC 14579)	✓	X	✓	✓	X	X	✓	✓	✓	X	✓
<i>B. clausii</i> (KSM-K16)	✓	X	✓	✓	X	X	✓	✓	X	✓	✓
<i>B. halodurans</i> (C-125)	✓	✓	✓	✓	X	✓	✓	✓	✓	X	✓
<i>B. licheniformis</i> (ATCC 14580)	✓	X	✓	✓	X	X	X	X	✓	X	✓
<i>B. subtilis</i> (strain 168)	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓
<i>B. thuringiensis</i> <i>konkukian</i> (strain 97-27)	✓	X	✓	✓	X	X	✓	✓	✓	X	✓

Table 5.2: Datasets used to construct the individual PFINs, where KEGG pathway data was the gold standard for each respective network.

of 0.2 allows new members to be added if their node scores are less than the seed node by a maximum of 20%). The smaller the cutoff, the more finegrained the clusters; by default, this is set to 0.2 [MCODE, <http://baderlab.org/Software/MCODE>]. Scores given by MCODE to the clusters are determined by multiplying the density of the cluster by the number of members/nodes [Bader and Hogue, 2003].

5.3 Results

5.3.1 Threshold determination for network optimisation

The *Bacillus* PFINs were analysed further in order to determine a threshold weight at which to remove weakly associated interactions by comparing the PFINs against the gold standard dataset, KEGG pathway. The results are illustrated in figure 5.1, showing the distribution of pairs based on their composite weights. Based on this graph, a composite cutoff weight of five was chosen for creating the refined *B. subtilis* PFIN. This decision was based on the number of TPs recovered and the size of the resulting PFIN (i.e. the total number of interactions). The plots of the other organisms show a similar trend to *B. subtilis*, so a common cutoff weight of five was also applied to these PFINs.

5.3.2 Network properties of *Bacillus* PFINs

A number of quantitative measures are frequently applied to networks:

- *Degree*

The degree refers to the connectivity of a node k i.e. how many edges the node has to other nodes. For an undirected network, with N nodes and E edges the average degree $\langle k \rangle$ is calculated as in equation 5.2.

$$\langle k \rangle = 2E/N \quad (5.2)$$

- *Degree distribution*

The degree distribution $P(k)$ gives the probability that a node has exactly k edges. $P(k)$ is calculated by counting the number of nodes $N(k)$ with k edges and dividing by the total number of nodes N (equation 5.3). The degree distribution may define the type of the network.

$$P(k) = N(k)/N \quad (5.3)$$

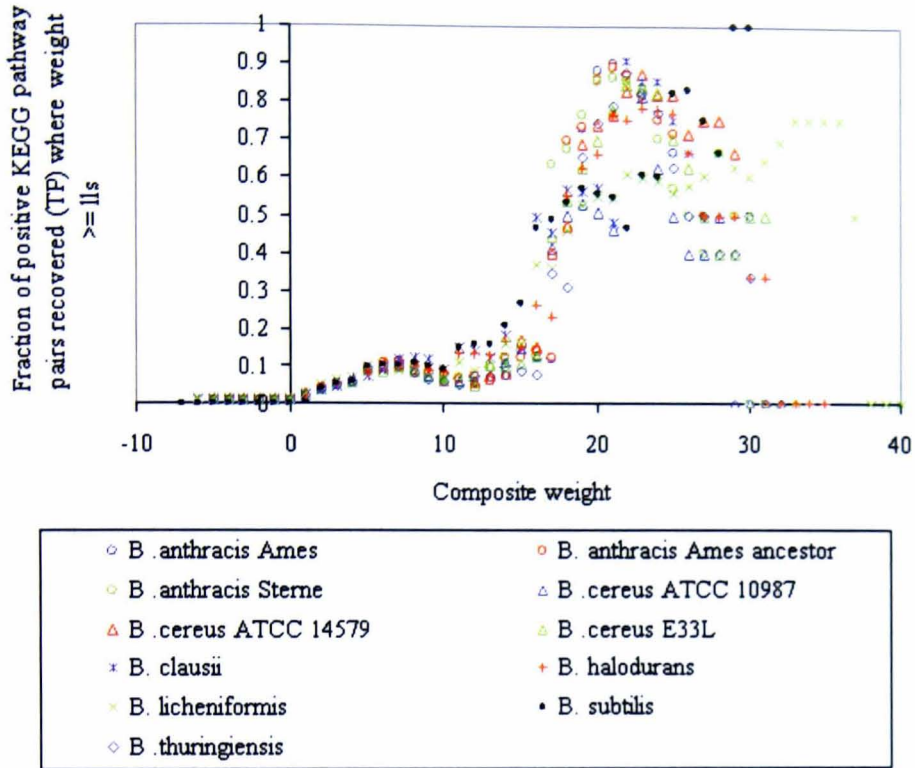


Figure 5.1: By comparison to KEGG pathway, the distribution of TPs, across the *Bacillus* PFINs, was measured as the composite weight increases.

- *Shortest path and average path length*

Distance in networks is measured using path length i.e. how many edges are passed in order to travel between two nodes. As there may be alternative paths between two nodes, the *shortest path* is used to measure distance. The average shortest path length between every possible pair of nodes is therefore used as a network measure.

- *Clustering coefficient*

The modularity of the network can be summarised by using a metric called the clustering coefficient. This is measured for node i using the clustering coefficient C_i (equation 5.4) based on the number of edges for this node k_i and the number

of edges connecting the k_i neighbours of the node i to each other, n_i . C_i therefore measures the number of triangles that go through node i . However, $k_i(k_i - 1)/2$ is the total number of triangles that could pass through node i , if all of its neighbours are connected to each other.

$$C_i = 2n_i/k_i(k_i - 1) \quad (5.4)$$

The average clustering coefficient measures the likelihood of nodes to form clusters. The average clustering coefficient of all nodes with k links is measured using $C(k)$. This function is important in defining the structure of the network [Assenov, 2006, Barabási and Oltvai, 2004, Li et al., 2006].

The network statistics for the 11 *Bacillus* PFINs were generated using a Cytoscape plugin, called NetworkAnalyzer [Assenov, 2006]. Observation of the basic network properties shown in table 5.3 reveals the *B. subtilis* PFIN is the least connected of all the networks, having the lowest number of edges and largest average shortest path length. *B. licheniformis* is the most highly connected, having the lowest average shortest path length, the least number of nodes, but containing an average number of edges. This probably reflects the broader number of evidence types that were incorporated in the *B. subtilis* PFIN, and consequently, the greater level of detail, than was used in developing the *B. licheniformis* PFIN.

The network properties captured by $P(k)$ and $C(k)$ are independent of the size of the network, therefore these functions define the common properties of the network and consequently they can be used to categorise the network. By analysing these measures, networks can be compared and characterised against each other [Barabási and Oltvai, 2004, Li et al., 2006].

The network distribution plots, shown in table E.1 indicate that the PFINs appear to be small-world networks. The networks small-world properties are characterised according to their clustering coefficients and average shortest path lengths. Small-world networks are characterised by an average shortest path length that is comparable to their equivalent randomly connected network. The distance between any two nodes is

Genomes	No. of nodes	No. of edges	Average no. of neighbours	Average shortest path length	Connected pairs
<i>B. anthracis</i> (Sterne)	2160	40486	37.487	11	4176132 (89%)
<i>B. anthracis</i> (Ames ancestor)	2030	37555	37.0	9	3765714 (91%)
<i>B. anthracis</i> (Ames)	2019	37082	36.733	10	3746304 (91%)
<i>B. cereus</i> (E33L)	2596	57241	44.099	11	5479602 (81%)
<i>B. cereus</i> (ATCC 10987)	2090	47088	45.060	10	3954396 (90%)
<i>B. cereus</i> (ATCC 14579)	2331	43625	37.430	10	4505978 (82%)
<i>B. clausii</i> (KSM-K16)	1955	45416	46.461	10	3260070 (85%)
<i>B. halodurans</i> (C-125)	2068	41723	40.351	9	3318964 (77%)
<i>B. licheniformis</i> (ATCC 14580)	1795	36001	40.113	8	2767440 (85%)
<i>B. subtilis</i> (strain 168)	2547	34791	27.319	15	4777066 (73%)
<i>B. thuringiensis konkukian</i> (strain 97-27)	2446	38588	31.552	11	4604736 (76%)

Table 5.3: Basic network statistics.

small despite the majority of nodes not being neighbours to one another. In addition, small-world networks have a larger proportion of neighbourhood connectivity, highlighted by having higher clustering coefficients than their equivalent random networks [Assenov, 2006, Barabási and Oltvai, 2004, Li et al., 2006, Watts and Strogatz, 1998]. For a random network, the clustering coefficient is equivalent to the edge density. Therefore, in order to determine whether the *Bacillus* networks are small-world, the clustering coefficients (*CC*) and average shortest path length (*ASP*) of the *Bacillus* networks and their respective random networks were compared by determining if the conditions stated in equations 5.5 and 5.6 were met. Both these conditions were true for each of the *Bacillus* PFINs (table 5.4). The maximum ratio for CCs between the real and the random benchmark networks was 98%, whereas the maximum ratio for ASPs between the real and the random benchmark networks was 28%.

$$\frac{CC_{real}}{CC_{random}} > 1.2 \quad (5.5)$$

$$\frac{ASP_{real}}{ASP_{random}} < 2 \quad (5.6)$$

Genomes	CC_{real}	CC_{rand}	ASP_{real}	ASP_{rand}
<i>B. anthracis</i> (Sterne)	0.556	0.017	3.199	2.506
<i>B. anthracis</i> (Ames ancestor)	0.554	0.017	3.195	2.501
<i>B. anthracis</i> (Ames)	0.545	0.017	3.201	2.506
<i>B. cereus</i> (E33L)	0.525	0.017	3.280	2.461
<i>B. cereus</i> (ATCC 10987)	0.550	0.021	3.255	2.359
<i>B. cereus</i> (ATCC 14579)	0.516	0.015	3.383	2.540
<i>B. clausii</i> (KSM-K16)	0.399	0.024	3.051	2.318
<i>B. halodurans</i> (C-125)	0.584	0.019	3.151	2.447
<i>B. licheniformis</i> (ATCC 14580)	0.543	0.022	3.012	2.395
<i>B. subtilis</i> (strain 168)	0.485	0.010	3.815	2.745
<i>B. thuringiensis konkukian</i> (strain 97-27)	0.542	0.012	3.439	2.655

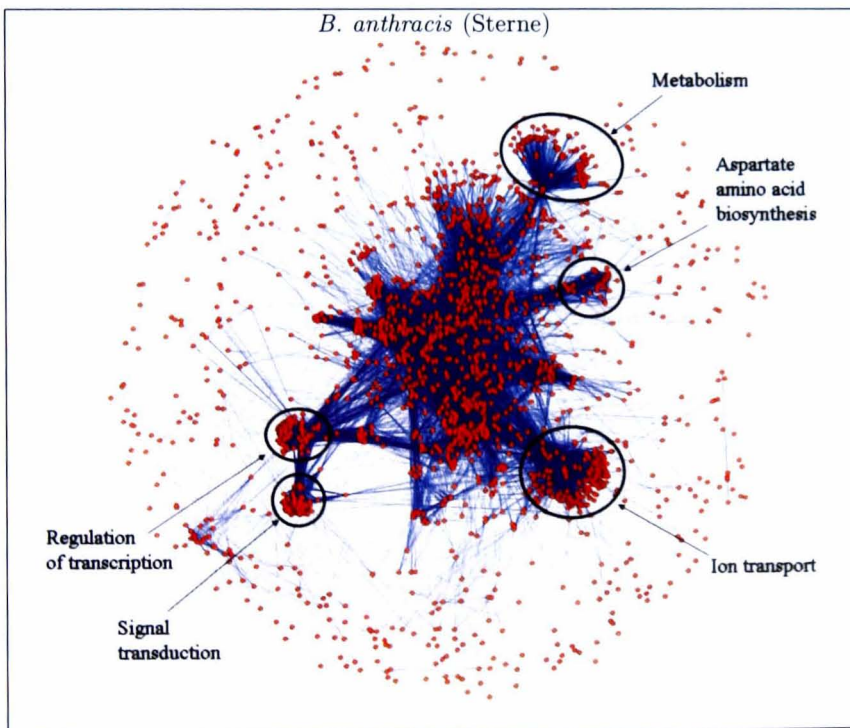
Table 5.4: Small-world network properties.

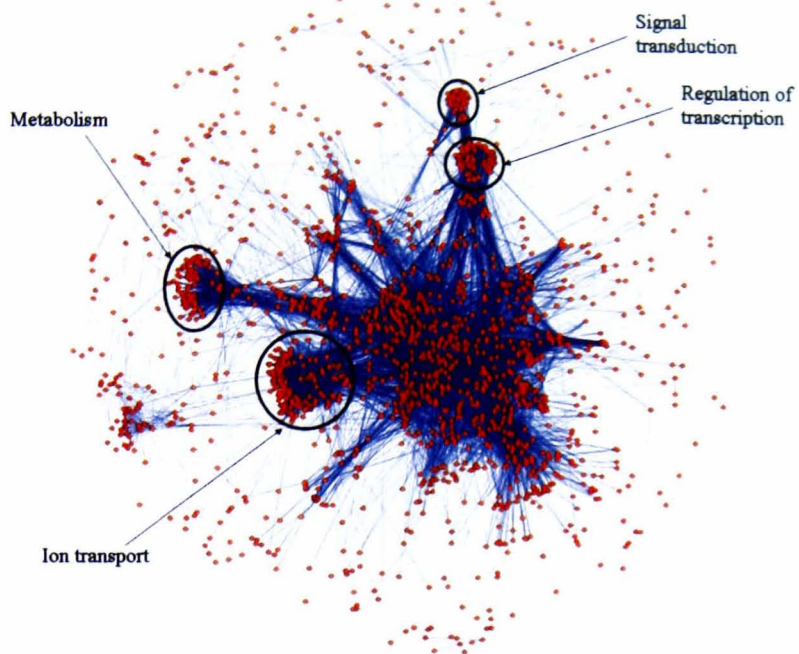
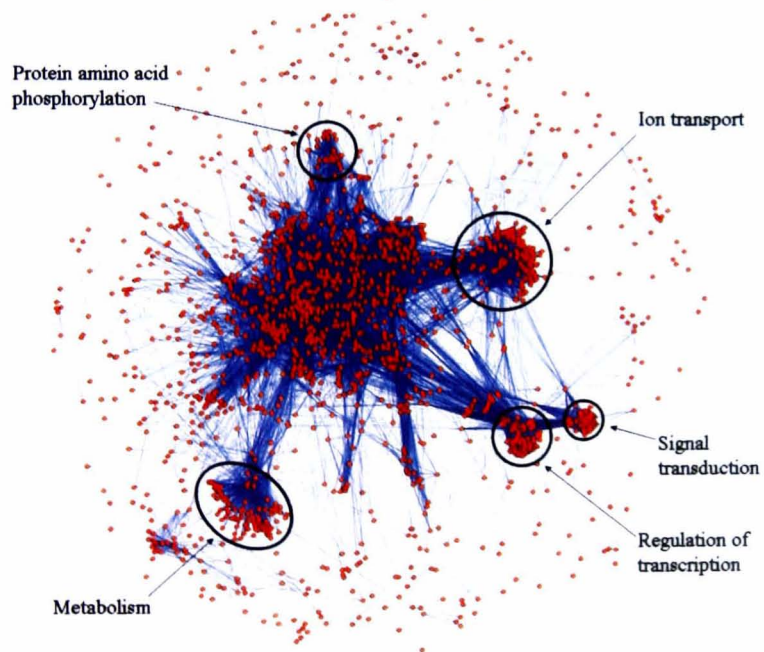
The networks were then analysed for scale-free properties. Scale-free networks consist of nodes with one or two links, to nodes with many links (hubs). The defining feature of these networks is a power law relation. The linear relation in the node degree distribution plots based on logarithmic scores highlights the potential scale-free nature of these networks. However, a statistical measure of the models goodness of fit is provided by calculating the coefficient of determination R^2 . It is generally accepted that for scale-free networks R^2 is greater than 0.8. The *Bacillus* networks therefore verge on being scale-free, with R^2 values ranging from 0.63 for *B. licheniformis* to 0.78 for *B. subtilis* (indicated in table E.1), compared to the respective random networks which range from 0.007 to 0.102.

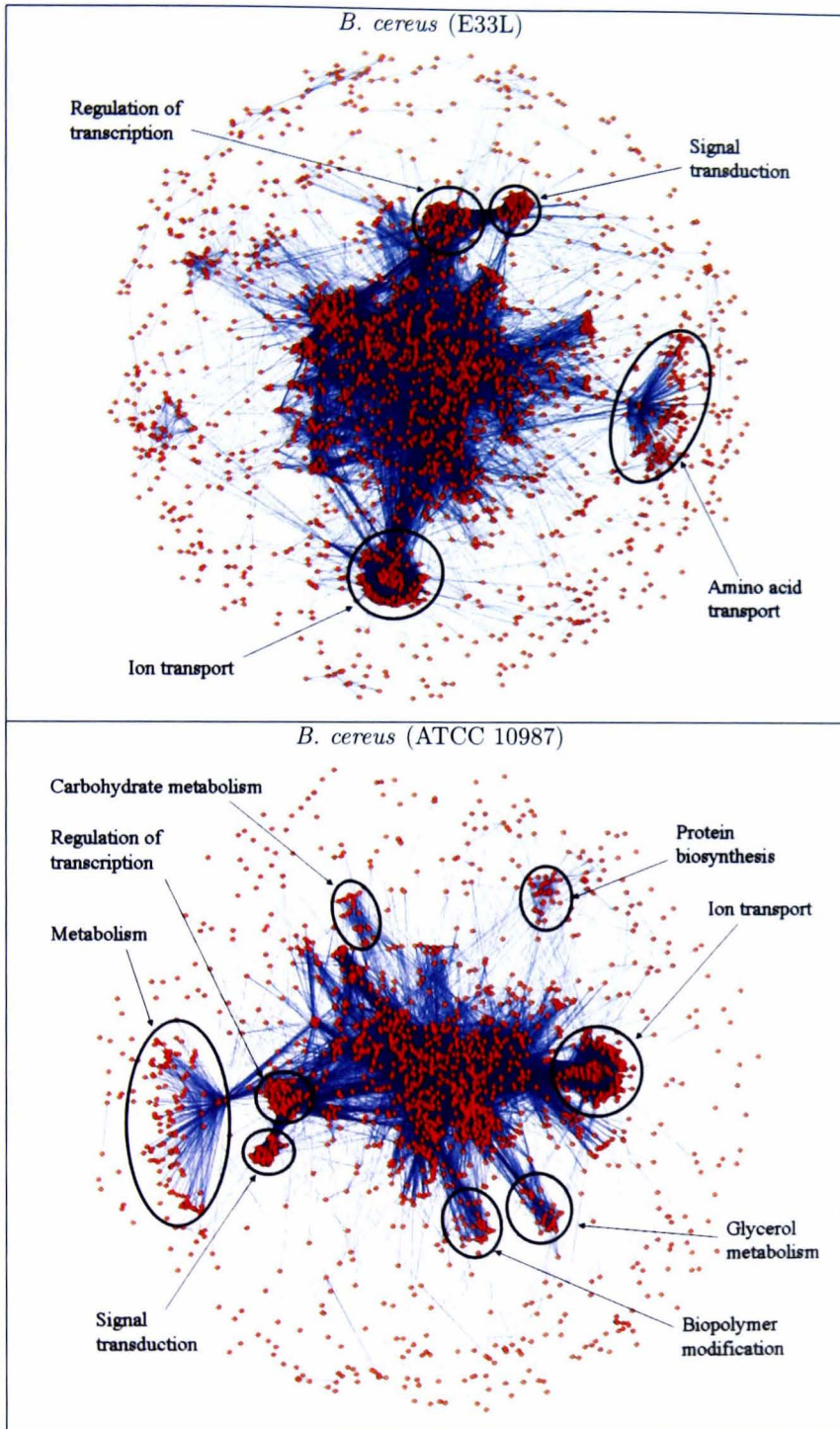
An alternative means of evaluating the PFINs is by graphically visualising and querying the networks in order to assess their 'correctness'. By manually delving into the protein interactions of the PFINs, implementation errors and weaknesses can be uncovered.

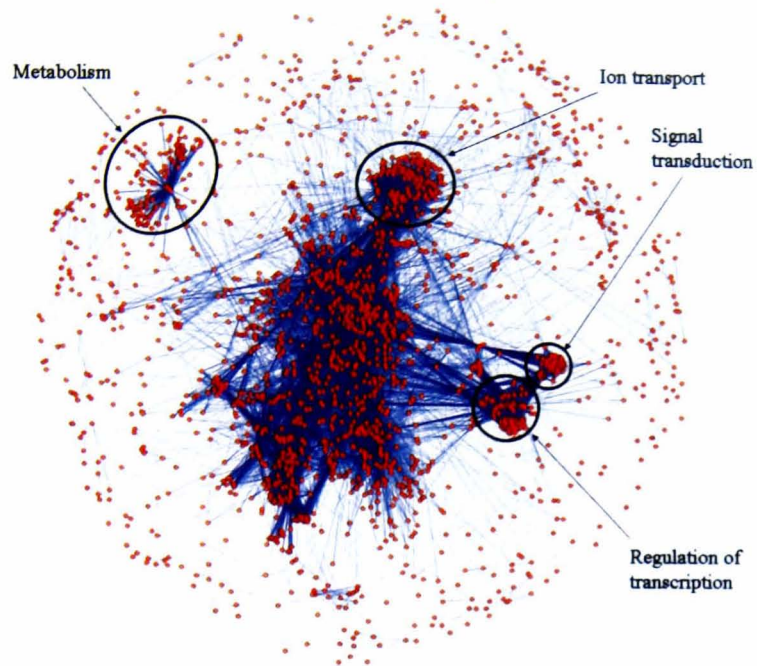
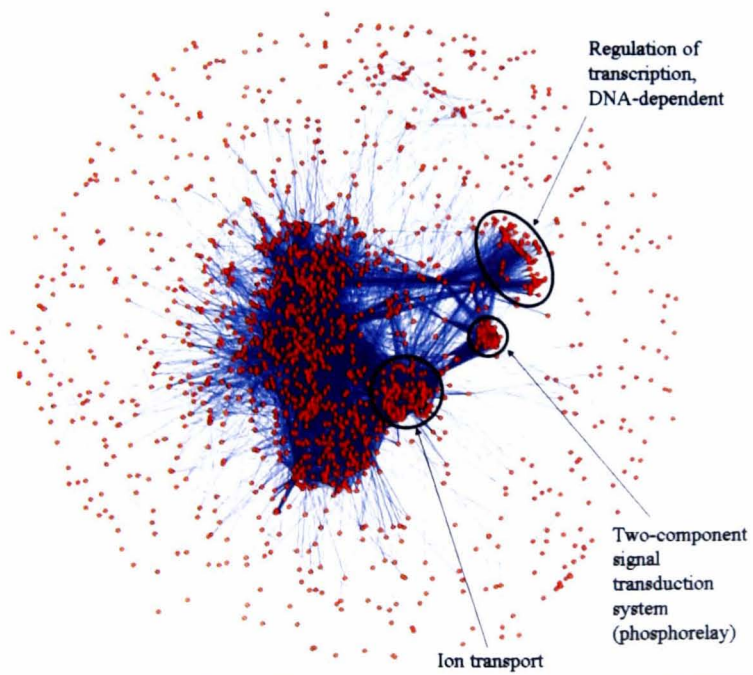
Visualisation of the PFINs developed through an edge-weighted spring embedded layout, using Cytoscape, highlights the existence of a number of clusters (table 5.5). Using the Cytoscape plugin, BiNGO, functional annotations could be associated with the major clusters based on GO process terms. Manual inspection reveals all 10 *Bacillus* PFINs possess clusters related to the regulation of transcription, signal transduction and transport. Other identifiable clusters relate to metabolism and biosynthesis. A unique cluster was found in the *B. halodurans* PFIN showing transposition behaviour. Also, in the *B. licheniformis* PFIN, a cluster responsible for cell redox homeostasis was identified.

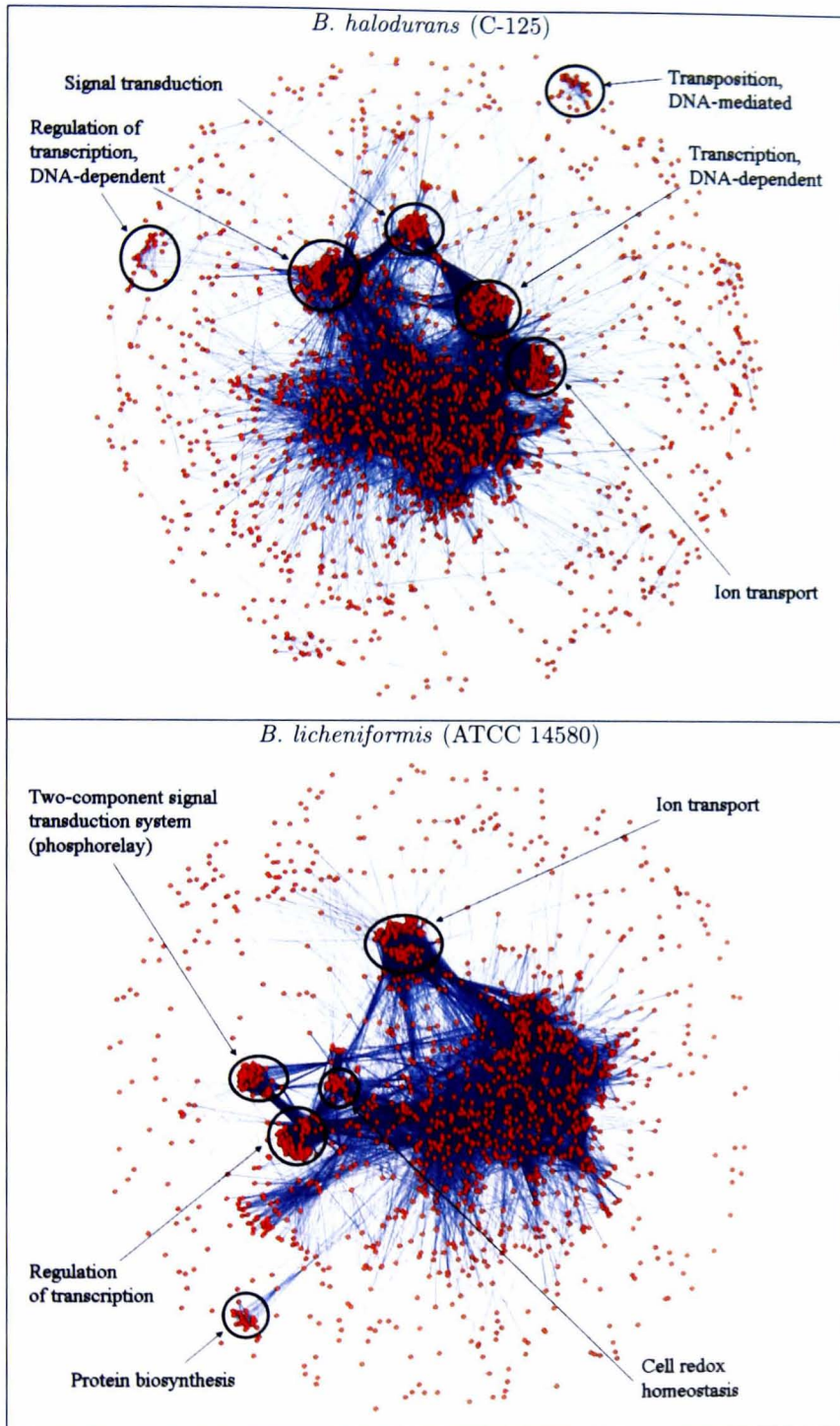
Table 5.5: PFINs for the 11 *Bacillus* species, as viewed using the edge-weighted spring embedded layout in Cytoscape. The main clusters detected visually are annotated using over-represented GO process terms.

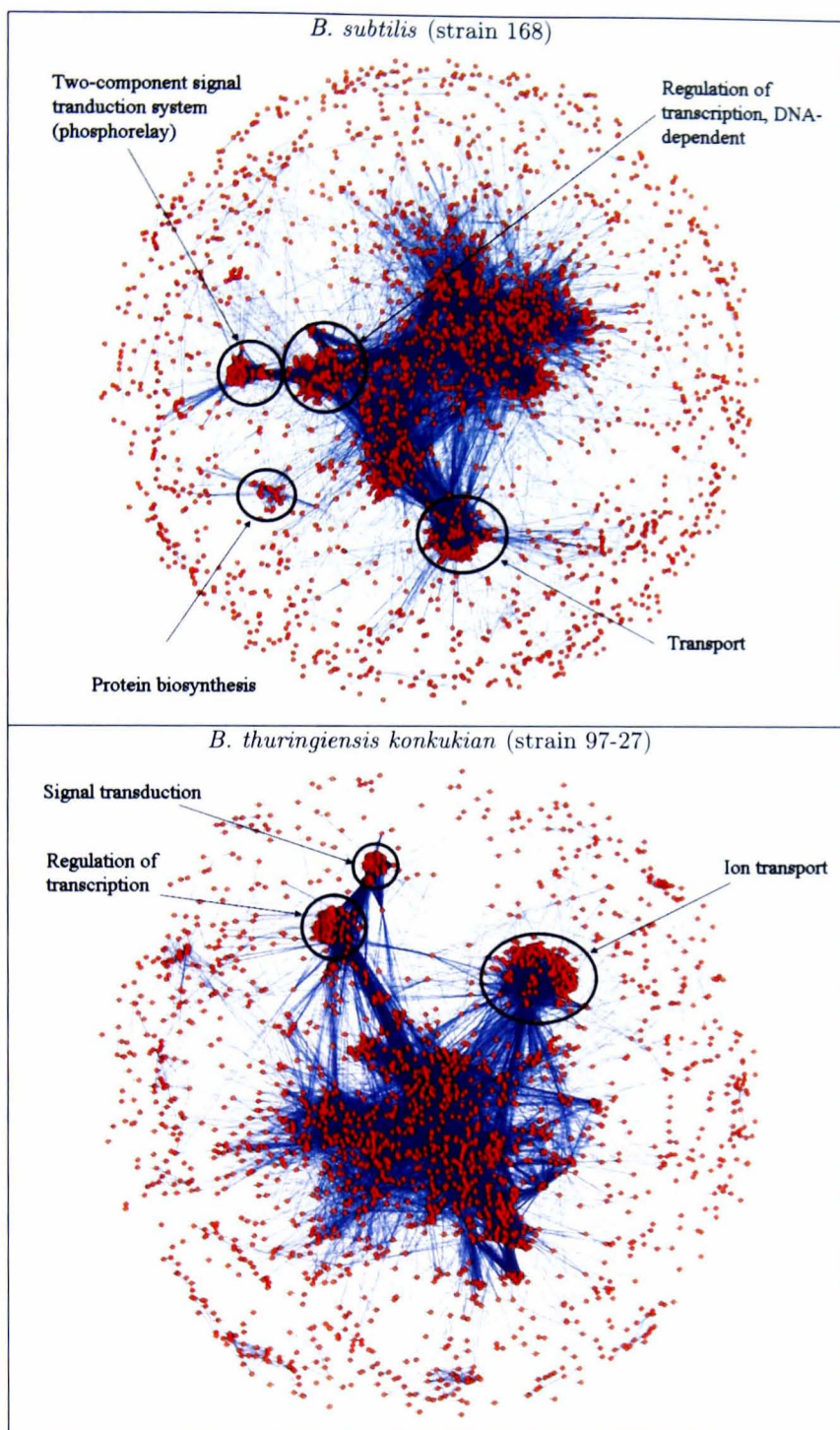


B. anthracis (Ames ancestor)*B. anthracis* (Ames)



B. cereus (ATCC 14579)*B. clausii* (KSM-K16)





5.3.3 Determination of optimal clustering settings of the *B. subtilis* PFIN

In order to assess the value most appropriate for clustering the interactions of the *B. subtilis* PFIN, a range of node cutoff scores (0.0, 0.1, 0.2, 0.3, 0.4, 0.5) were used to cluster the *B. subtilis* PFIN. The average composition of the clusters produced were investigated with regard to the mean and modal number of members per cluster and the MCODE score per cluster, as well as the denseness of the clusters, assessed by summing the MCODE scores. The plot used to decide which score to use as a cutoff is shown in figure 5.2. In this graph the node cutoff score is plotted against the sum of all the cluster scores for the respective node cutoff score. It was concluded the optimal node cutoff score for clustering the *B. subtilis* PFIN was 0.1, due to the peak in the graph at this point.

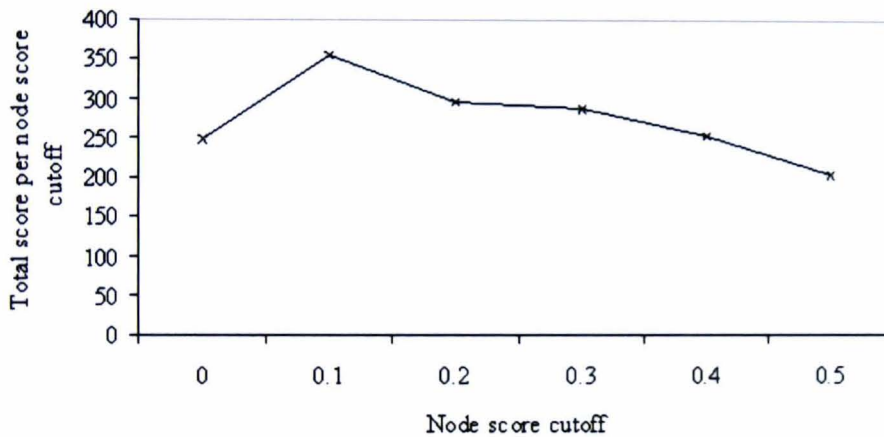


Figure 5.2: Granularity of clusters based on different node cutoff scores.

The highest scoring MCODE clusters were then reclustered using a node cutoff value of 0.0. However, the newly formed clusters were not significantly different from the original clusters. Reclustering resulted in the formation of only one new cluster that is more tightly clustered. For instance, the top cluster of *B. subtilis*, which after initial clustering using parameter 0.1 contains 99 nodes, was reclustered using a

parameter 0.0 resulted in 3 clusters, containing 51, 5 and 6 nodes each. Reclustering the second cluster returns only one cluster, reduced in size from 59 to 38 nodes. The third cluster is also reduced, from 41 to 32 nodes. It was therefore decided to continue using the original set of clusters, without loss of data.

To visualise the clusters an additional Cytoscape plugin called GenePro [Vlasblom et al., 2006] was used. Using this plugin, clusters are shown as nodes and edges represent inter-cluster protein interactions. The variation in cluster size is reflected in the size of the nodes. GenePro also allows the incorporation of node attribute information by displaying a node (cluster) as a pie chart.

Following the determination of the most appropriate node cutoff score for the *B. subtilis* PFIN, the other *Bacillus* PFINS were clustered using the same value, equal to 0.1. The cluster information was stored in the database allowing simple querying over all the clusters across all species.

5.3.4 Functional analysis of the clustered *B. subtilis* PFIN

A number of distinguishable clusters have previously been highlighted in the *Bacillus* PFINS (figure 5.5). However, the application of an automatic clustering algorithm, such as MCODE, enables the identification of less obvious clusters.

An overview of the MCODE-identified clusters in the *B. subtilis* PFIN, and their functional breakdown based instead on SubtiList classification codes (due to the more appropriate number of categories), is shown in figure 5.4. A more detailed view of the major clusters identified in the *B. subtilis* PFIN, using MCODE, is shown in table F.1.

5.3.5 Distribution of secreted proteins within the topology of the *B. subtilis* PFIN

As shown in the case of *B. subtilis*, the predicted secreted proteins identified by BaSPP are fairly distributed throughout the PFIN (figure 5.3). This distribution can be shown in terms of the functional modules/clusters identified in the *B. sub-*

tilis PFIN, reflecting the global spread of the putative secreted proteins across the many functionally interacting clusters (figure 5.5). The functionally distributed nature of the secretome is further illustrated in figure 5.4 in which the cluster member functions are summarised in terms of SubtiList classification codes. A more detailed view of these clusters is provided in table F.1 in which both functional and secreted information is combined to give an overall view of each cluster.

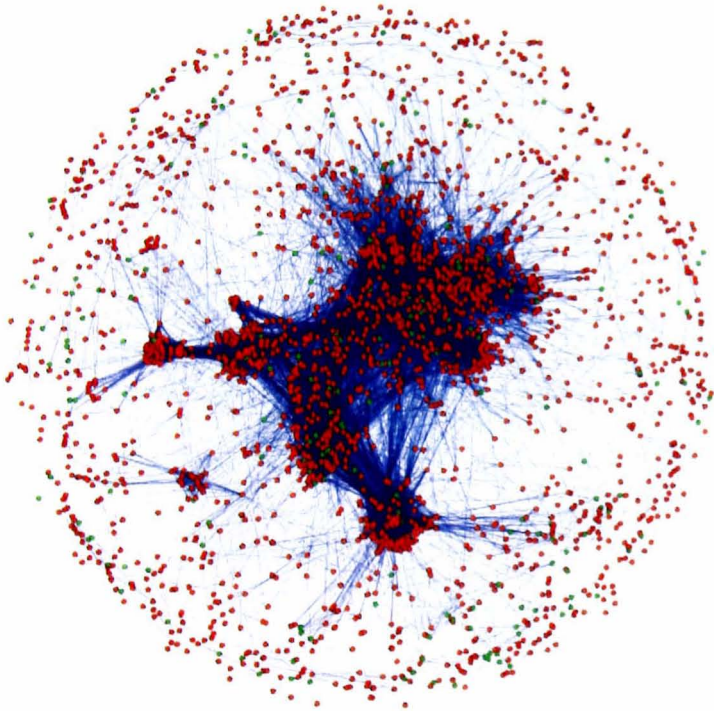


Figure 5.3: *B. subtilis* PFIN. Green dots represent the putative secreted proteins and the red dots are the non-secreted proteins.

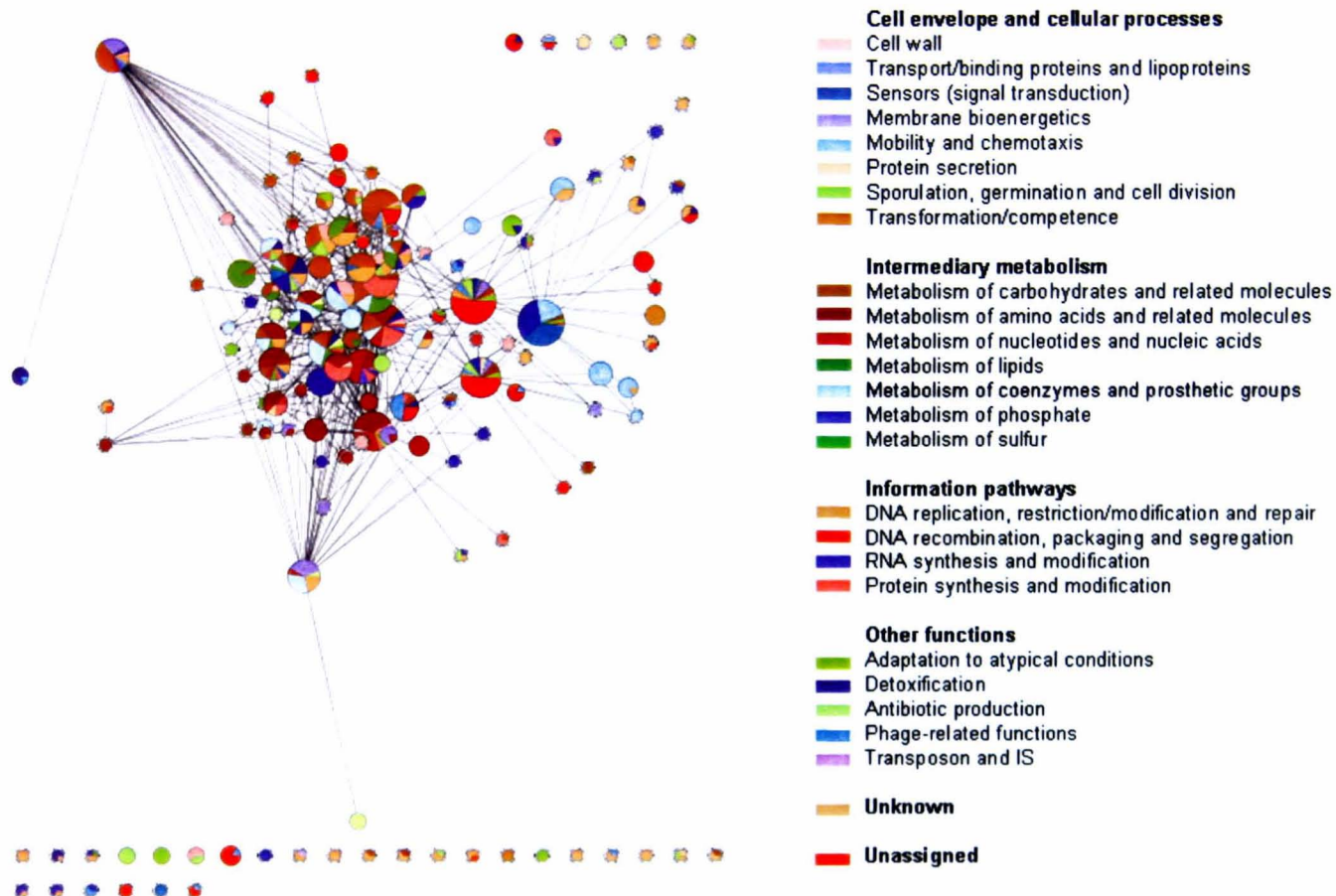


Figure 5.4: Functional distribution of the highly connected clusters in the *B. subtilis* (strain 168) PFIN, according to SubtiList, where a pie chart represents the functional breakdown of a cluster and edges represent interactions between 1 or more proteins in 2 clusters. The size of the pie charts reflects the size of the clusters.

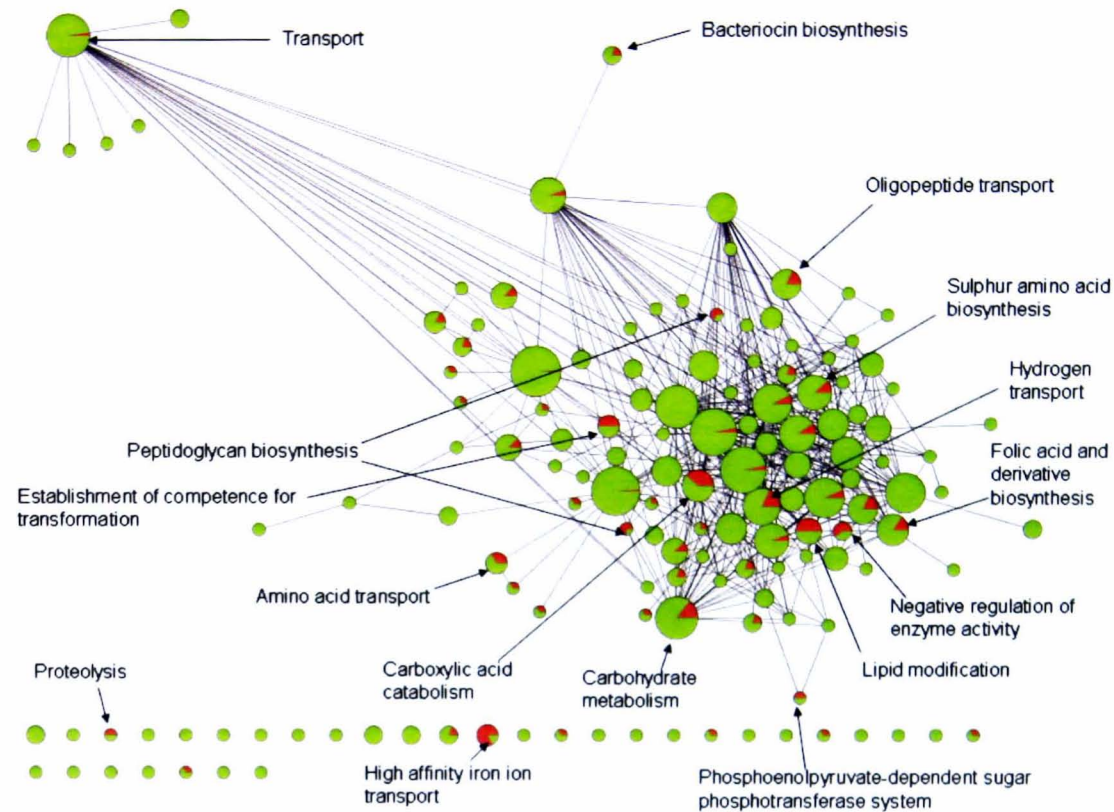


Figure 5.5: Distribution of putative secreted proteins, as predicted by BaSPP, through the highly connected clusters of the *B. subtilis* (strain 168) PFIN. Each circle/pie chart represents a cluster, where red wedges represent secreted proteins and the green represent non-secreted proteins. The size of the pie charts reflects the size of the clusters. A few of the clusters containing a proportion of secreted proteins are annotated with over-represented GO process terms for that cluster.

5.4 Discussion

PFINs were constructed for each of the *Bacillus* species in this study. A data warehouse has been developed containing all the information gathered on each of the *Bacillus* PFINs. The core database schema is not just specific to *Bacillus* species, but can be applied to capture the PFIN information of any species.

The capability of the *B. subtilis* PFIN to recover protein interactions identified in the KEGG pathway database is shown in figure 5.1. However, within this figure, a common trend can be seen for the other *Bacillus* species. After a composite pair weight is exceeded, the number of TPs decrease, unlike that observed for *B. subtilis*. This suggests that KEGG, or the datasets themselves, are incomplete, and a lot still remains unknown about these other *Bacillus* species.

The lack of data sharing and accessibility were factors that limited the integration of a number of resources into the networks. There were cases where appropriate data sources were found that could be incorporated into the PFIN, but the inability to download these resources prevented this from happening e.g. SPiD [Hoebeker et al., 2001].

For those data sources that were used, in some cases these could have been used more effectively. The method of incorporating BLAST results is one such example, which could have been improved by using Inparanoid [Remm et al., 2001] to generate orthologous protein links. Another example is the method of inferring protein pairs based on GO. A more thorough approach to including GO predicted protein-protein interactions is to estimate the semantic similarity between annotated proteins [Lord et al., 2003]. This approach was employed in interactome development in Brown and Jurisica [2005] and Wu et al. [2006].

The integration of a variety of data sources to build the PFINs was important in ensuring a large degree of coverage in terms of the size of the proteome, as the coverage of each experimental dataset differs. Each experiment has different associated error rates too [Lee et al., 2004]. Therefore using data integration methods strengthens links that are common.

Relying on a gold standard has many drawbacks in itself. It prevents the inclusion of the gold standard dataset into the networks. Perhaps more significantly, the final weightings associated with each link between a pair of genes is dependent on the accuracy and coverage of the gold standard. The development of methods to integrate data sources without having to rely on a gold standard are therefore needed; such a method has been documented in Deng et al. [2004].

The design and implementation of SubtilNet allows for easy expansion, whether that be to include additional evidence datasets or to extend to other species. The core framework is generic and therefore does not require any modification in order to use.

The majority of the tasks performed in computing a PFIN are computationally intensive. Therefore, a greater utilisation of e-science technology could be most beneficial. The current system setup is e-science-based in the sense that system makes use of shared data that is integrated in order to calculate the PFINs. However, transferring the current 'single host system setup' to a distributed workflow-based implementation would not only improve usability, but also provides a means of using remote resources to perform intensive calculations.

Developing these PFINs provides a mechanism to explore unknown functions and hypotheses that would otherwise not be so readily accessible. Specifically, data concerning unknown and previously unidentified links in the PFINs can be determined, from which functional predictions and relations can be judged. Functions can be inferred by viewing and querying the PFINs individually, but also via cross-species comparison.

Chapter 6

The application of PFINs to the systems level analysis of secreted protein families

6.1 Introduction

This chapter documents cross-species interactome comparison between *Bacillus* species. Interactions between members of the protein families representing similar secreted *Bacillus* proteins, identified by BaSPP, were investigated.

The analysis of these PFINs has been made easier with the availability of computational methods capable of reducing and visualising interactomes, either within a single interactome or between interactomes, enabling a more efficient means of extracting information e.g. by using clusters. Interactome comparison is an area of active research, with still much scope for improvement. This is particularly true in regards to prokaryotes. Analysing the proteins in terms of their interactions within and across species can lead to further insights into the properties of the secretomes of *Bacillus* species.

6.2 Methodology

A number of secreted protein families, previously identified as interesting by the BaSPP system, were further explored in the context of their PFINs. Each family of

interest was taken separately. For each family, clusters identified by MCODE that were found to contain a protein member were extracted. In cases where proteins did not belong to a cluster, first-neighbours were used instead.

Links between the separate species-specific clusters (or first-neighbours) were then inferred using orthology. To identify orthologous interactions the program Inparanoid was used [Remm et al., 2001]. This program has the benefit of providing confidence scores based on BLAST, which can be used to weight the vertical cross-species links. Paralogues were ignored in this process, with only the main orthologue pairs used to infer cross-species links. As orthologue links are weighted differently from links within the PFINs, the scores associated with the horizontal and vertical links were normalised by dividing the edge weights by the maximum edge weights, for the PFIN and orthologous links respectively. The links were then integrated to provide a cross-species subnetwork representation of the regions of interest. This process resulted in a multi-species sub-PFIN consisting of horizontal interactions (within the same species) and vertical interactions (across species).

The cross-species subnetwork of highly connected clusters were then displayed in Cytoscape, in order to visualise the interactions and their meaning. The network was colour-coordinated to highlight the proteins belonging to different organisms. All sub-networks were then analysed using a Cytoscape plugin, called BiNGO [Maere et al., 2005], to identify those GO categories that were over-represented in the subnetwork.

6.3 Results

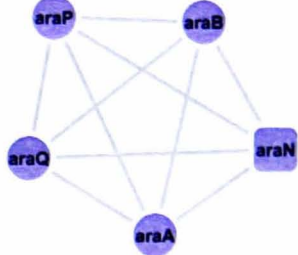
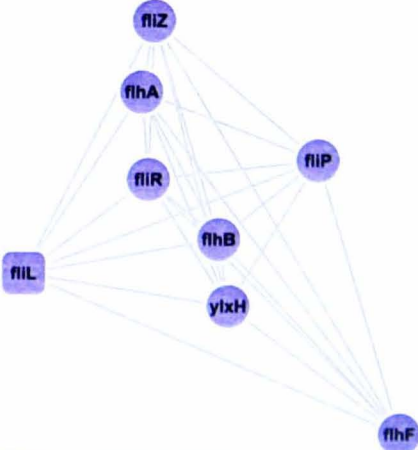
6.3.1 Analysis of cross-species clusters incorporating protein families

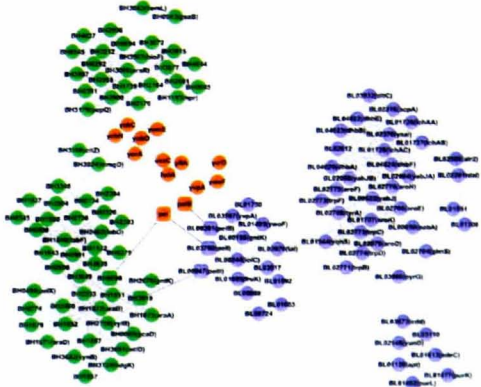
Nine protein families identified in chapter 4 were investigated in terms of their interactions, in order to identify additional information to that gathered by sequence analysis techniques. This corresponded to five out of nine of the non-pathogen-specific families, three out of ten of the pathogen-specific families, and the core PrsA family. In some

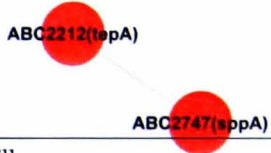
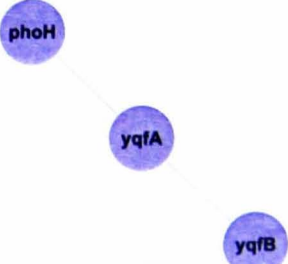
instances, members of the protein families are not represented within their respective PFINs, either because the composite weights associated with their interactions are all less than the threshold cutoff, which is generally the case, or because there are no associated interactions involving the protein, according to the evidence datasets used. This explains the absence of certain members of the remaining protein families from being represented in their protein family cross-species subnetworks, but also explains why the remaining four non-pathogen-specific families and seven pathogen-specific families were not analysed. The core PrsA family was also analysed as much still remains unknown about the function of PrsA and its interactions, despite being well understood in terms of its structure. The known pathogen and non-pathogen specific families were also the focus of the analysis. The cross-species subnetworks, or in some cases only a species-specific subnetwork, of the pathogen- and non-pathogen-specific families, as well as the core protein family PrsA, are shown in table 6.1.

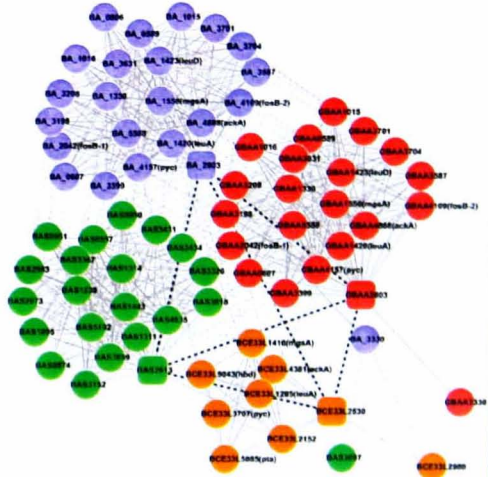
Table 6.1: The cross-species subnetworks representing the protein families (identified by BaSPP) specific to the known pathogens and non-pathogens, as well as the core protein family encoded by *prsA*. Apart from *B. subtilis*, whose members are represented by gene names, the members belonging to the other species are represented as locus tags, with the gene name in brackets if known. *B. licheniformis* (strain ATCC 14580) refers to the substrain Novozymes for which a PFIN was developed, not Goettingen. Three letter abbreviations of the *Bacillus* species are used to define the colour coding. Square boxes indicate family members.

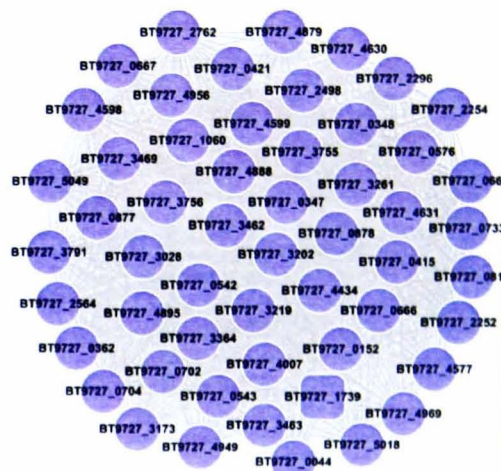
Family	Protein members	Cross-species subnetwork
CORE FAMILY		
<i>prsA</i>	<p>Family members:</p> <p>BC_1043 - Protein export protein <i>prsA</i> precursor</p> <p>BC_1161 - Peptidyl-prolyl cis-trans isomerase</p> <p>BC_2272 - Protein export protein <i>prsA</i> precursor</p> <p>BC_2862 - Protein export protein <i>prsA</i> precursor</p> <p>BCE33L0952 - protein export protein</p> <p>BCE33L1061 - peptidyl-prolyl cis-trans isomerase</p> <p>BCE33L2101 - protein export protein</p> <p>BL00855 - putative PpiC-type peptidyl-prolyl cis-trans isomerase</p> <p><i>yacD</i> - similar to protein secretion PrsA homologue</p> <p><i>prsA</i> - molecular chaperone</p> <p>BT9727_0962 - protein export protein <i>prsA</i></p> <p>BT9727_1066 - peptidyl-prolyl cis-trans isomerase</p> <p>BT9727_2115 - protein export protein <i>prsA</i></p> <p>BT9727_2619 - peptidylprolyl isomerase</p> <p>(Note: GBAA1041, GBAA1169, GBAA2336, BA_1041, BA_1169, BA_2336, BAS0974, BAS1084, BAS2178, BCE_1145, BCE_1280, BCE_2359, ABC1527, BH1177 and BL02827 are absent from refined PFINs.)</p> <p>Subnetwork members:</p> <p>BL00853(<i>yacB</i>), BL00854(<i>hslO</i>), BL00857(<i>cysK</i>), <i>map</i>, <i>yacC</i>, <i>clpX</i>, <i>clpP</i>, <i>htrA</i>, BCE33L0220(<i>murF</i>), BCE33L4438(<i>murC</i>), BCE33L3673(<i>murE</i>), BCE33L3500(<i>pepT</i>), BCE33L3671(<i>murD</i>), BCE33L1980(<i>murF</i>)</p>	<p>Blue = <i>bsu</i>, green = <i>bce</i>, orange = <i>bcz</i>, red = <i>bli</i>, yellow = <i>bt</i>k</p>

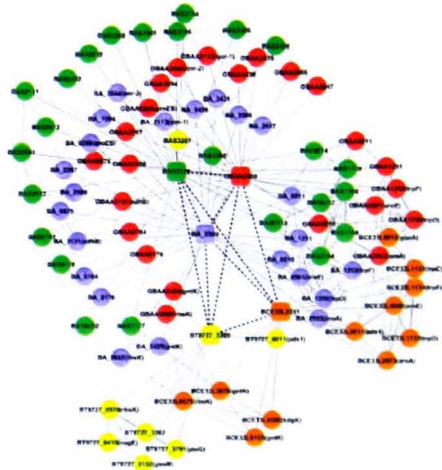
Family	Protein members	Cross-species subnetwork
NON-PATHOGENIC FAMILIES		
<i>araN</i>	<p>Family members: <i>araN</i> - sugar-binding protein (Note: ABC3471, ABC3774, ABC0385, ABC3301, ABC3215, BH0905, BH1117 and BL00348 are absent from the refined PFINs.)</p> <p>Subnetwork members: <i>araP</i>, <i>araQ</i>, <i>araA</i>, <i>araB</i></p>	<p>Blue = bsu</p> 
<i>fliL</i>	<p>Family members: <i>fliL</i> (Note: ABC2259, BH2447 and BL01263 are absent from the refined PFINs.)</p> <p>Subnetwork members: <i>fliP</i>, <i>fliA</i>, <i>fliB</i>, <i>fliZ</i>, <i>ylxH</i>, <i>fliF</i>, <i>fliR</i></p>	<p>Blue = bsu</p> 

Family	Protein members	Cross-species subnetwork
<i>pelA</i> / <i>pelB</i>	<p>Family members:</p> <p>BH0698 - pectate lyase</p> <p>BH3819 - high-alkaline pectate lyase</p> <p>BL00947 - pectate lyase, Polysaccharide Lyase Family 1</p> <p>BL03760 - pectate lyase family 1, PelI</p> <p>BL00361 - pectate lyase, Polysaccharide Lyase Family 1</p> <p><i>pel</i> - pectate lyase</p> <p><i>pelB</i> - pectate lyase</p> <p>(Note: ABC0063 is absent from the refined PFIN.)</p> <p>Subnetwork members:</p> <p>BH1739, BH3072, BH2996, BH0292, BH1952, BH1873(<i>araA</i>), BH1872(<i>araB</i>), BH3460, BH2676(<i>gntK</i>), BH2708, BH3649, BH2061, BH0981, BH1937, BH3348, BH1880, BH3333, BH0275, BH3815, BH0391, BH0774, BH0545, BH3077, BH1878, BH2384, BH2504, BH2696, BH3635, BH4044, BH3179(<i>pepQ</i>), BH0694, BH1551, BH1871(<i>araD</i>), BH2756(<i>xykB</i>), BH4037, BH3887, BH1840(<i>fabF</i>), BH1843, BH0494(<i>pelX</i>), BH2756(<i>xykB</i>), BH4037, BH3887, BH1840(<i>fabF</i>), BH1843, BH0494(<i>pelX</i>), BH3851(<i>mtlD</i>), BH1185(<i>hpr</i>), BH3212, BH3683(<i>xynB</i>), BH1132, BH3907(<i>bioF</i>), BH2800, BH2492(<i>fabD</i>), BH2164, BH0945, BH0943(<i>gsaB</i>), BH3160(<i>citZ</i>), BH3000(<i>arsR</i>), BH2734, BH0080, BH0326, BH3724(<i>kdgK</i>), BH0645, BH2303, BH3924(<i>mmgD</i>), BH2176, BH1867, BH0065(<i>gcaD</i>), BH1857, BH3043(<i>hemL</i>), BH2086, BL04023(<i>dhbB</i>), BL02612, BL02064(<i>yabJA</i>), BL02316(<i>acpA</i>), BL02079(<i>aroD</i>), BL04022(<i>dhbE</i>), BL02370(<i>yaaI</i>), BL01727(<i>lchAB</i>), BL02201(<i>dal</i>), BL02771(<i>trpB</i>), BL03970(<i>tal</i>), BL02774(<i>trpD</i>), BL03966(<i>pyrG</i>), BL02566(<i>alr2</i>), BL01728(<i>lchAC</i>), BL02776(<i>aroH</i>), BL01707(<i>aroK</i>), BL04020(<i>dhbA</i>), BL03932(<i>dltC</i>), BL03110, BL01726(<i>lchAA</i>), BL02779(<i>aroF</i>), BL00724, BL00522(<i>yabJ</i>), BL00859(<i>pabA</i>), BL01544(<i>yqhS</i>), BL02773(<i>trpC</i>), BL02772(<i>trpF</i>), BL01482(<i>purL</i>), BL01126(<i>apt</i>), BL01892, BL00244(<i>iolC</i>), BL02768(<i>tyrA</i>), BL01477(<i>purK</i>), BL02766(<i>aroE</i>), BL02068(<i>yabJB</i>), BL03517, BL01495(<i>ywoF</i>), BL03673(<i>cdd</i>), BL01605(<i>fruK</i>), BL00869, BL01851, BL04024(<i>dhbF</i>), BL02148(<i>yunD</i>), BL01308, BL01613(<i>adeC</i>), BL00195(<i>gntK</i>), BL01663, BL03597(<i>yvpA</i>), BL02704(<i>glmS</i>), BL01750, <i>yomE</i>, <i>yfiA</i>, <i>hxlA</i>, <i>uraC</i>, <i>yclG</i>, <i>yvpA</i>, <i>ywoF</i>, <i>yobO</i>, <i>ycbC</i>, <i>yorA</i>, <i>ycbH</i></p>	<p>(See figure G.4 for larger version.)</p> <p>Blue = ban, red = bar, green = bat, orange = bcz, yellow = btk.</p> 

Family	Protein members	Cross-species subnetwork
<i>sppA</i>	Family members: ABC2747 - signal peptidase SppA (Note: BH3198, BL00425 and <i>sppA</i> are absent from the refined PFINs.)	Red = bcl 
	Subnetwork members: ABC2212(<i>tepA</i>)	
<i>yqfA</i>	Family members: <i>yqfA</i> - unknown (Note: ABC1672, BH1357 and BL01411 are absent from the refined PFINs.)	Blue = bsu 
	Subnetwork members: <i>phoH</i> , <i>yqfB</i>	

Family	Protein members	Cross-species subnetwork
PATHOGENIC FAMILIES		
A	<p>Family members: GBAA2803 - S-layer protein, putative BA2803 - S-layer protein, putative BAS2613 - S-layer protein, putative BCE33L2530 - S-layer protein (Note: BT9727_2561 is absent from refined PFIN.)</p> <p>Subnetwork members: BA0607, BA1420(<i>leuA</i>), BA0806, BA1330, BA1015, BA3704, BA3631, BA4157(<i>pyc</i>), BA1423(<i>leuD</i>), BA1016, BA1556(<i>mgsA</i>), BA3208, BA5588, BA0589, BA4109(<i>fosB-2</i>), BA2042(<i>fosB-1</i>), BA3587, BA4888(<i>ackA</i>), BA3701, BA3198, BA3399, BA3330, GBAA1330, GBAA4109(<i>fosB-2</i>), GBAA4888(<i>ackA</i>), GBAA4157(<i>pyc</i>), GBAA1420(<i>leuA</i>), GBAA1016, GBAA5588, GBAA3631, GBAA3399, GBAA3587, GBAA0589, GBAA1423(<i>leuD</i>), GBAA0607, GBAA3701, GBAA1556(<i>mgsA</i>), GBAA2042(<i>fosB-1</i>), GBAA3208, GBAA3704, GBAA3198, GBAA3330, GBAA1015, BAS1314, BAS0574, BAS3326, BAS3859, BAS3367, BAS3818, BAS5192, BAS3431, BAS1230, BAS0557, BAS3152, BAS1311, BAS4535, BAS2973, BAS1443, BAS2983, BAS0951, BAS0950, BAS1895, BAS3434, BAS3087, BCE33L4381(<i>ackA</i>), BCE33L2530, BCE33L1285(<i>leuA</i>), BCE33L5043(<i>hbd</i>), BCE33L5085(<i>pta</i>), BCE33L2980, BCE33L2152, BCE33L1416(<i>mgsA</i>), BCE33L3707(<i>pyc</i>)</p>	<p>(See figure G.2 for larger version.)</p> <p>Blue = ban, red = bar, green = bat, orange = bcz, yellow = btk.</p> 

Family	Protein members	Cross-species subnetwork
E	<p>Family members: BT9727_1739 - conserved hypothetical protein (Note: GBAA1900, BA1900, BAS1762 and BCE33L1711 are absent from refined PFINs.)</p> <p>Subnetwork members: BT9727_0704, BT9727_0733(<i>scrA</i>), BT9727_0152(<i>glmM</i>), BT9727_1060, BT9727_0421, BT9727_3469(<i>chiA</i>), BT9727_0878(<i>bglH</i>), BT9727_0347, BT9727_0576(<i>rbsK</i>), BT9727_3364, BT9727_2296(<i>iolC</i>), BT9727_0702(<i>celB</i>), BT9727_0542(<i>treB</i>), BT9727_5049(<i>pfoR</i>), BT9727_0348(<i>ptsG</i>), BT9727_3755(<i>nplT</i>), BT9727_0415(<i>nagE</i>), BT9727_0543(<i>treC</i>), BT9727_0818, BT9727_2254(<i>celB</i>), BT9727_0044(<i>gcaD</i>), BT9727_0877(<i>bglP</i>), BT9727_0362(<i>chiA</i>), BT9727_0666(<i>scrK</i>), BT9727_0667(<i>scrB</i>), BT9727_0668(<i>scrA</i>), BT9727_2252(<i>ynpK</i>), BT9727_2498(<i>pulA</i>), BT9727_2564(<i>ampD</i>), BT9727_2762, BT9727_3028, BT9727_3173, BT9727_3202, BT9727_3219, BT9727_3261(<i>amyS</i>), BT9727_3462(<i>fruA</i>), BT9727_3463(<i>fruB</i>), BT9727_3756(<i>malL</i>), BT9727_3791(<i>ptsG</i>), BT9727_4007, BT9727_4434(<i>amyX</i>), BT9727_4577(<i>kdgK</i>), BT9727_4598(<i>glgC</i>), BT9727_4599(<i>glgB</i>), BT9727_4630(<i>gtaB</i>), BT9727_4631(<i>manB</i>), BT9727_4879(<i>ugd</i>), BT9727_4888(<i>celB</i>), BT9727_4895(<i>celB</i>), BT9727_4949, BT9727_4956(<i>gtaB</i>), BT9727_4969(<i>murA</i>), BT9727_5018(<i>murA2</i>)</p>	<p>(See figure G.1 for larger version.)</p> <p>Blue = btk</p> 

Family	Protein members	Cross-species subnetwork
L	<p>Family members:</p> <p>GBAA2500 - hypothetical protein BA2500 - hypothetical protein BAS2320 - hypothetical protein BCE33L2241 - group-specific protein BT9727_2285 - hypothetical protein</p> <p>Subnetwork members:</p> <p>BA0784, BA2953(<i>aroA</i>), BA1251, BA0011, BA3131(<i>adhB</i>), BA3544(<i>qor-2</i>), BA1250(<i>trpD</i>), BA0010, BA3435, BA1252(<i>trpF</i>), BA0675, BA3428(<i>gntK</i>), BA0665(<i>rhsK</i>), BA1004, BA2647, BA2267, BA3438, BA2113(<i>qor-1</i>), BA0266(<i>groES</i>), BA4561(<i>aroE</i>), BA0176, BA2588, BA3566, GBAA2953(<i>aroA</i>), GBAA2647, GBAA1250(<i>trpD</i>), GBAA2267, GBAA3428(<i>gntK</i>), GBAA0665(<i>rhsK</i>), GBAA0010, GBAA1252(<i>trpF</i>), GBAA0266(<i>groES</i>), GBAA1251, GBAA3131(<i>adhB</i>), GBAA2113(<i>qor-1</i>), GBAA2588, GBAA1004, GBAA4561(<i>aroE</i>), GBAA0675, GBAA0784, GBAA3566, GBAA0176, GBAA0011, GBAA3438, GBAA3544(<i>qor-2</i>), GBAA3435, BAS0641, BAS2912, BAS1160, BAS2744, BAS2111, BAS1158, BAS0013, BAS1159, BAS3306, BAS3177, BAS3286, BAS3186, BAS4232, BAS0939, BAS0747, BAS2466, BAS0014, BAS0632, BAS3184, BAS2412, BAS0178, BAS3260, BAS3261, BAS1965, BAS0252, BCE33L1133(<i>trpC</i>), BCE33L0012(<i>guaA</i>), BCE33L4080(<i>aroE</i>), BCE33L1134(<i>trpF</i>), BCE33L0011(<i>pdxI</i>), BCE33L1132(<i>trpD</i>), BCE33L0575(<i>rhsK</i>), BCE33L2673(<i>aroA</i>), BCE33L4596(<i>kdgK</i>), BCE33L0155(<i>gntK</i>), BCE33L3078(<i>gntK</i>), BT9727_3202, BT9727_0415(<i>nagE</i>), BT9727_0152(<i>glmM</i>), BT9727_3791(<i>ptsG</i>), BT9727_0011(<i>pdxI</i>), BT9727_0576(<i>rhsK</i>)</p>	<p>(See figure G.3 for larger version.)</p> <p>Blue = ban, red = bar, green = bat, orange = bcz, yellow = btk.</p> 

6.3.2 A detailed cross-species analysis of the core PrsA protein family and interacting partners using PFINs

Cross-species PFINs were used to investigate the distribution of PrsA orthologues and interacting functional partners within the core PrsA protein family. The objective of the study was to demonstrate how the PFINs can be used to shed light on the function of the secreted protein families by focusing on the PrsA protein family as a use case.

6.3.2.1 The PFIN for the core PrsA protein family

The PrsA cross-species subnetwork, shown in table 6.1, did not provide a detailed picture of the functional interactions in which PrsA is involved, as not all PrsA family members are represented. Proteins belonging to *B. anthracis* (Ames), *B. anthracis* (Ames ancestor), *B. anthracis* (Sterne), *B. cereus* (ATCC 10987), *B. clausii*, *B. halodurans* and *B. licheniformis* were not included in the PrsA cross-species PFINs. This was because these proteins were not present in the respective PFINs, as the composite weight associated with each of their functional interactions were less than the threshold (set to five). It was subsequently decided this threshold was simply too high. To overcome this, a lower threshold of two was used to construct a new set of PFINs for each *Bacillus* species. As expected, these new PFINs contain each member of the PrsA protein family. Visual analysis of orthologous relations within the PrsA family reveals that there are three distinct sub-families of *prsA* orthologues (figure 6.1).

B. subtilis possesses a single copy of the *prsA* gene [Harwood pers. comm] which falls into a subcluster together with a single copy of a *prsA* orthologue from each of the other 10 species. In this analysis *B. clausii* and *B. halodurans* also appear to have one copy of *prsA*. All of the *B. anthracis* strains and *B. cereus* (E33L) have three *prsA* homologues each, whilst *B. cereus* (ATCC 14579) and *B. thuringiensis* appear to possess four *prsA* homologues.

The products of the three single copy *prsA* homologues from *B. subtilis*, *B. clausii*

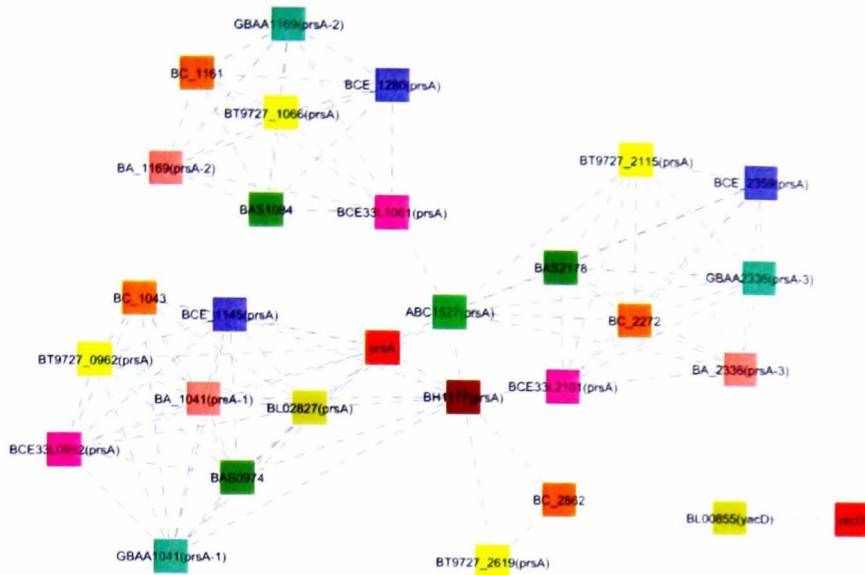


Figure 6.1: A graph-based view of the orthologous relationships in the PrsA protein family. Dark green = *B. anthracis* (Sterne), light blue = *B. anthracis* (Ames ancestor), light pink = *B. anthracis* (Ames), dark pink = *B. cereus* (E33L), dark blue = *B. cereus* (ATCC 10987), orange = *B. cereus* (ATCC 14579), light green = *B. clausii* (KSM-K16), brown = *B. halodurans* (C-125), pale green = *B. licheniformis* (ATCC 14580, sub_strain Novozymes), red = *B. subtilis* (strain 168) and yellow = *B. thuringiensis konkukian* (strain 97-27).

and *B. licheniformis* appear to be central to the network, implying that they may have been the root from which the multiple copies of *prsA* present in *B. anthracis*, *B. cereus* and *B. thuringiensis* have evolved. Two additional *prsA* copies from *B. thuringiensis* and *B. cereus* (ATCC 14579) appear to be most similar to the single *prsA* copy from *B. halodurans*.

Examination of the annotation from the EMBL records of these genes reveals considerable variation in annotation. Of the 28 proteins in the family, half are annotated with the term 'export protein *prsA*', whilst the others have various annotations suggesting that they are peptidyl-prolyl cis-trans isomerases (PPIases), enzymes that are responsible for helping proteins fold correctly. PrsA is a membrane-bound lipoprotein that is known to help proteins fold correctly during the secretory process [Wahlström

et al., 2003] and has a domain that is homologous with parvulin type PPIases, which explains this annotation and the orthologous relationships within this family.

One of the main uses of PFINs are in the prediction of function for genes which have not yet been characterised, using the guilt-by-association approach [Aravind, 2000, Oliver, 2000]. One candidate for such prediction is present in the *prsA* homology subnetwork. In addition to those proteins in the family annotated with the term PPIase, *yacD* is a protein from *B. subtilis* that is of unknown function and not yet annotated. The association of *yacD* with PPIases in this family through orthology alone, allows us to hypothesise that, it too, is a PPIase; a hypothesis which would be easily testable within the laboratory. Similar predictions can be made with respect to other protein families.

The value of the PFIN approach is not, however, restricted to the inference of protein function. Other discrepancies with published knowledge can arise, particularly since many of the datasets used to construct a PFIN are derived from publicly-available, but unpublished, databases. Consider, for example, the number of homologues of *prsA* identified in the different *Bacillus* species using this approach. To date, there has been no reported systematic experimental analysis of *prsA* and its copy number in different *Bacillus* species. However, a recent experimental analysis of *B. anthracis* reveals that it actually possesses four functional homologues of PrsA [Williams, 2002], whilst our PFIN identifies only three in all three *B. anthracis* strains investigated. We used quite stringent cutoffs for the identification of homologous genes, a factor which will undoubtedly affect the number of genes identified. The relationship between strength of homology, as measured by Inparanoid scores, and gene function in different species clearly requires further investigation. Analyses of this kind further illustrates the value of PFINs as hypothesis-generation tools in biology, and specifically the cross-species *Bacillus* PFIN for integrating and directing research about this fascinating and economically-valuable group of organisms.

6.3.2.2 The global structure of PFIN for the core PrsA protein family and functional interacting partners

Whilst an examination of the orthologous relationships in the cross-species PFINs is a valuable approach to help assign function to uncharacterised proteins, another powerful way is by visual and computational analysis of the functional relationships between the proteins in the *Bacillus* PFINs within individual species.

The cross-species PFIN for the PrsA protein family is shown in full in figure 6.2. Functional edges were cutoff at a composite weight of two and orthology was assigned between pairs of proteins whose sequences possessed reciprocal best BLASTp hits, using the Inparanoid program. The resulting network contains 367 nodes representing proteins, and 4461 edges representing functional interaction between the proteins.

Figure 6.2 shows the members of the PrsA protein family (square nodes) and those proteins that are predicted to be closely functionally linked (round nodes). Whilst the size of this network precludes displaying full details of the nodes, the general structure of the network, including the extent of cross-species orthologous links (dotted lines) and the density of within-species functional edges (solid lines) is clear. The protein nodes fall into 11 clusters, each of which corresponds to a particular *Bacillus* species (as indicated with the colour coding in the figure). Notably, *B. subtilis* has the most dense cluster, almost certainly reflecting the more extensive amount of data available for this well-studied organism.

When applying this network to the inference of function for previously uncharacterised proteins, an individual protein may be selected and its immediate neighbourhood examined manually. However, the complexity of this network makes an exhaustive manual analysis unfeasible. Alternatively, computational analysis can be applied to the network. Algorithms for filtering out specific network properties, for identifying pathways, or for identifying structure, such as clusters, can be applied. Often a combination of these two strategies is optimal. In the next section such an analysis of the PrsA protein family is described, as an example of the type of analysis which can be performed.

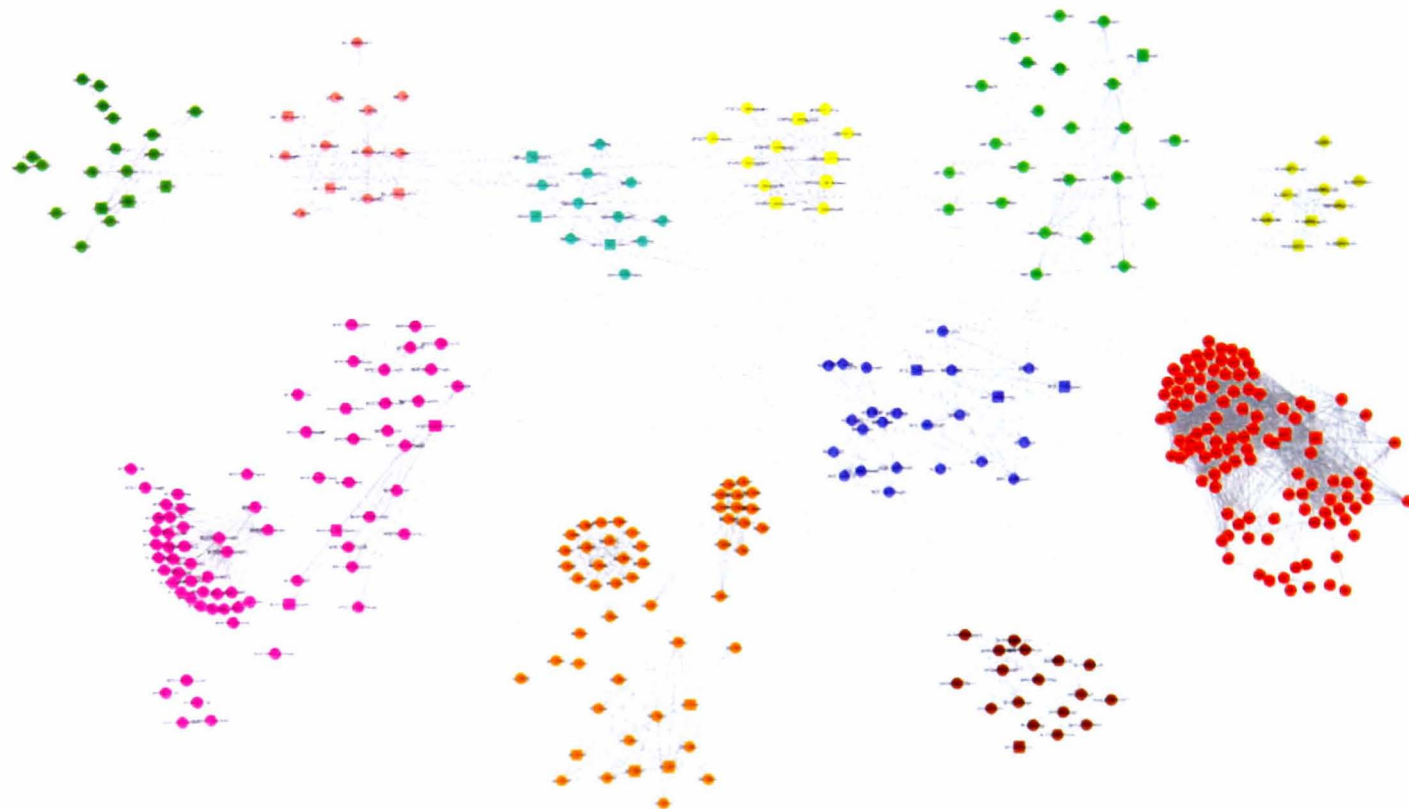


Figure 6.2: The cross-species PFIN for the PrsA protein family. Dark green = *B. anthracis* (Sterne), light blue = *B. anthracis* (Ames ancestor), light pink = *B. anthracis* (Ames), dark pink = *B. cereus* (E33L), dark blue = *B. cereus* (ATCC 10987), orange = *B. cereus* (ATCC 14579), light green = *B. clausii* (KSM-K16), brown = *B. halodurans* (C-125), pale green = *B. licheniformis* (ATCC 14580, sub_strain Novozymes), red = *B. subtilis* (strain 168) and yellow = *B. thuringiensis konkukian* (strain 97-27).

6.3.2.3 Cluster analysis of the PrsA protein family PFIN

Cluster analysis of the network (in figure 6.2) identifies a number of distinct clusters. Using the default values applied with the MCODE algorithm, approximately 30 clusters are produced. Interestingly, many of the clusters contain only nodes from a single species. Each individual cluster can be the focus of a systematic, manual, study. However, the cluster containing PrsA from *B. subtilis* is shown as an example. The distribution of this cluster in the cross-species PFIN is shown in figure 6.3 and its composition in figure 6.4. The majority of the edges in the cluster are assigned through co-expression. Extracting those nodes that share a co-expression level above 0.3 from this *B. subtilis* PrsA cluster reveals two co-expression networks showing that *prsA* is co-expressed with other known heat shock stress response genes such as the sigmaB-regulated genes *clpP*, *clpX*, *groEL* and *groES* [Hecker and Völker, 1998]. Interestingly, the gene *ywrO*, currently of unknown function, is a member of this co-expression subnetwork. *ywrO* does not possess any orthologous links in the overall network but is known to possess orthologues in other *Bacillus* species [Anlezark et al., 2002]. The known stress response proteins Ctc and Gsi also cluster tightly with a number of genes of unknown function, that are also potential targets for further directed studies.

The cluster that contains a protein of unknown function, *yacD* from *B. subtilis*, also serves as an example for further investigation. The distribution of this cluster in the cross-species PFIN is shown in figure 6.5 and its composition in figure 6.6. YacD appears to cluster with a number of ribosomal proteins, as well as protein folding proteins (Tig and PpiB).

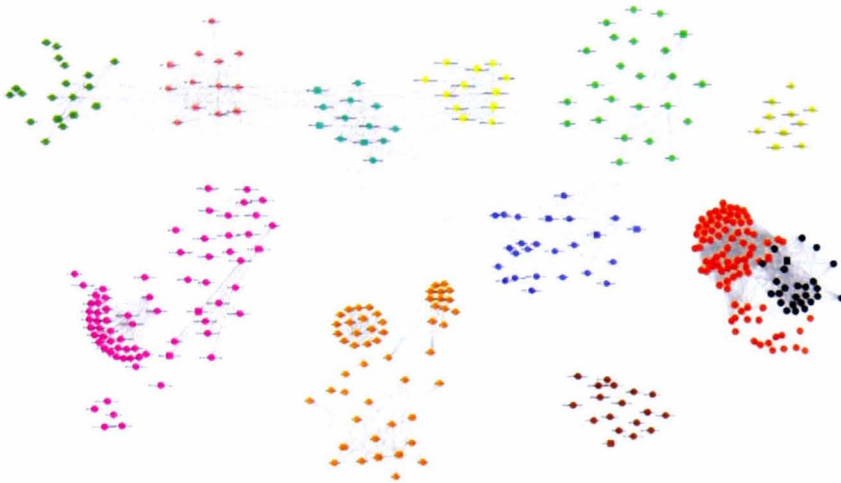


Figure 6.3: Distribution of the *prsA* cluster in the cross-species PFIN (shown in black).

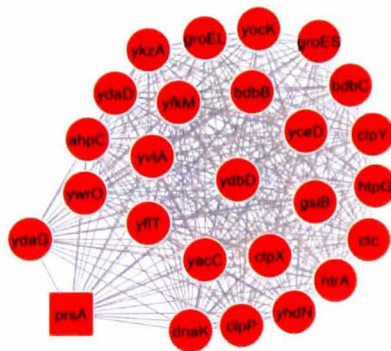


Figure 6.4: A close-up of the functional cluster derived from the PrsA family PFIN, containing *prsA* from *B. subtilis*.

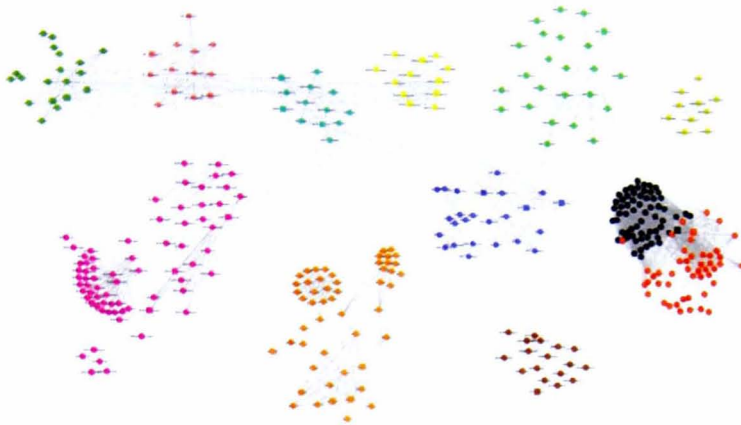


Figure 6.5: Distribution of the *yacD* cluster in the cross-species PFIN (shown in black).

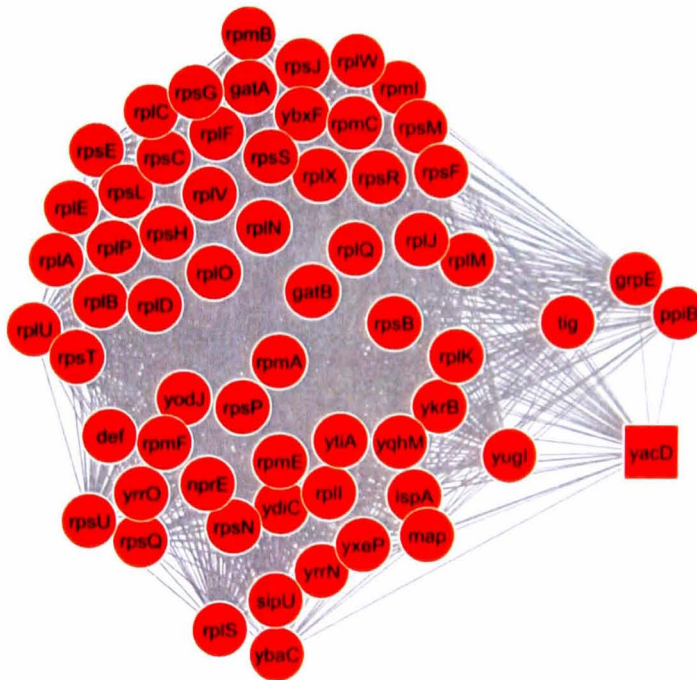


Figure 6.6: A close-up of the functional cluster derived from the PrsA family PFIN, containing *yacD* from *B. subtilis*.

6.4 Discussion

PFIN analysis is becoming a standard approach in systems biology. Analysis can infer the functions of uncharacterised proteins, or identify clusters and crosstalk between pathways (e.g. Hallinan and Wipat [2006]). To date, most analyses have been performed on single-species networks, or have involved the comparison of separate networks. Here, this approach has been extended to incorporate homology links between species to construct a single PFIN for multiple species of *Bacillus*. Special emphasis was paid to those interactions related to the secretomes of the *Bacillus* species. Specifically the pathogen and non-pathogen specific protein families uncovered in chapter 4, along with a core lipoprotein family, PrsA.

The majority of the non-pathogenic families previously identified in section 4.2.3.2 are not represented in the PFINs developed, as their interactions are below the threshold weight applied. Of the few that do appear in the PFINs, only the proteins AraN (BSU28750) and FliL (BSU16300), as well the PelA/PelB family, were found in highly connected regions (identified using MCODE).

Nothing substantial could be gained from the analysis of the arabinose utilisation proteins in *B. subtilis* than had previously been documented in Inácio et al. [2003] and Sá-Nogueira et al. [1997]. The arabinose utilisation pathway is used by the soil-based *Bacillus* species, specifically the non-pathogens, to degrade plant material [Inácio et al., 2003]. The cross-species PFIN is based mainly on operon and co-expression evidence data. It contains tightly clustered interactions involving other similar proteins (AraA, AraB, AraN, AraP, AraQ). Missing from this subnetwork are the other proteins involved in arabinose utilisation, and located on the same *ara* operon encoded by *araD*, *araE*, *araL*, *araM*, *araR* and *abfA*, which would have been expected to be observed in relation to these proteins. The only one of these proteins not present in cross-species PFIN but that is observed in the *B. subtilis* PFIN is AraE. The localisation of the AbfA enzyme is unknown, however predictions from BaSPP indicate this protein is not secreted and is localised in the cytoplasm. As an aside note, in addition to AraN, the other two secreted enzymes AraM and AraL are

predicted by BaSPP to be secreted and cytoplasmic, respectively [Inácio et al., 2003].

The flagellar network uncovered in the *B. subtilis* PFIN contains the non-pathogen-specific FliL protein, whose exact function is unknown. This cross-species PFIN contains other related proteins (FliZ, FliP, FlhA, FliR, FlhB, FlhF, YlxH) involved in flagellar synthesis. The interactions within the subnetwork are based on operon and co-expression data. It also appears, from Gram-negative studies, that this subnetwork mainly represents the flagellar export apparatus, thought to be composed of six transmembrane components FlhA, FlhB, FliO, FliP, FliQ and FliR, the three chaperones FliJ, FliS and FliT, the ATPase FliI, and its regulator FliH [Arnam et al., 2004]. Within the FliL subnetwork, four transmembrane components were identified (FlhA, FlhB, FliP and FliR). It has been suggested on the basis of the location of the *fliZ* gene in the *fliZPQR* cluster that in fact FliZ can be assigned as FliO. This would leave FliQ as being unassigned to this PFIN cluster. However, as FliQ appears in the overall PFIN with links between FlhA, FlhB, FlhF, FliL, FliM, FliR, FliY, this suggests it is simply a casualty of the clustering algorithm. As for FlhF, the function of this protein is unclear, although a GTP-binding motif has been identified. The final protein of this cluster is YlxH, another functionally unknown protein. It has been suggested this protein acts as a chaperone for translocation of flagellar proteins [Sonenshein et al., 2002].

Pectin lyases are a type of pectin-degrading enzymes secreted by many saprophytic and pathogenic organisms. Pectin-degrading enzymes are important industrially, for instance they have uses in the food and textile industries. Being able to understand their mechanisms and functions will help improve these industrial systems [Soriano et al., 2006]; studying the pectate lyase interactions in the PFINs can aid in this process. However, despite the identification of highly connected pectate lyase (PelA and PelB) regions in the *B. halodurans*, *B. licheniformis* and *B. subtilis* PFINs, with exception of *B. clausii*, there were minimal cross-species interactions identified based on orthology. This implies these proteins interact quite differently in the different species, which is further suggested by the different sizes of the organism specific connected regions. However, if the *B. subtilis* subnetwork is considered the most

accurate, it could also be suggested that there could be missing links in the Pel subnetworks of *B. halodurans* and *B. licheniformis* that could be inferred on the basis of the *B. subtilis* segment. The large proportion of uncharacterised proteins within the PelA and PelB regions also suggests a lot still remains unknown about pectate lyase. However, an interesting finding was the discovery of the two protein of unknown function YclG and YwoF, common across *B. subtilis* and *B. licheniformis*. The identification of these proteins within this subnetwork indicates that these two proteins are involved in the same pathway as PelA and PelB. However, YclG and YwoF are likely to function in quite different ways to each other. Just like PelA and PelB, YwoF is predicted to be secreted (as has been documented in Antelmann et al. [2001]) whereas YclG is localised in the cytosol.

The members of the signal peptidase *sppA* family can only be considered in terms the *B. clausii* PFIN. As this SppA protein does not lie within a highly connected region, consideration of its neighbours reveals it interacts with only one other protein, another SppA homologue, TepA. These proteins are isolated in the PFIN, with neither having any further interactions. The interaction is rather weak, based on STRING textmining data and GOA. The identification of this isolated interaction however provides no further indication about the functionality of these proteins than what has already been determined in Bolhuis et al. [1999]. However, both SppA and TepA are involved in secretion. TepA functions in the initial preprotein translocation stage of secretion, located in the cytoplasm (consistent with BaSPP's prediction), whereas SppA is a typically a membrane protein involved in the processing of preproteins [Bolhuis et al., 1999] (inconsistent with BaSPP). This latter case suggests the prediction made by BaSPP that SppA is secreted is erroneous. The lack of any additional transmembrane domains being predicted by TMHMM is a potential downfall, despite MEMSAT providing a positive transmembrane prediction.

Investigation of the YqfA protein family focused on the *B. subtilis* PFIN. Within the *B. subtilis* PFIN, YqfA neighbours a functionally unknown protein YqfB and the phosphate starvation induced ATPase PhoH. These three proteins (YqfA, YqfB and PhoH) exist in isolation. The interactions are based mainly on KEGG expression

data. This may suggest that the unknowns have functions similar to ATPases.

Of the 10 known pathogen-specific families identified in section 4.2.3.3, very few of the proteins were found in the PFINs developed (by thresholding at a weight of five). However, of the proteins in the pathogenic family E, BT9727_1739 was identified in a highly connected region. Analysis of this cluster reveals a high proportion of proteins involved in the carbohydrate transport system, phosphotransferase system (PTS). Due to the secretory nature of this protein and the proportion of related proteins involved in carbohydrate transport, this implies BT9727_1739 is a polysaccharide-degrading enzyme, secreted into the environment to allow the organism to reduce the polysaccharides into smaller components for the uptake and consequent metabolism of these entities [Sonenshein et al., 2002]. The predictive involvement in carbohydrate transport is supported by the finding in section 4.2.3.3 that this protein is likely a glycosyl hydrolase, BNR repeat-containing protein. It is therefore likely that this family of proteins, including BT9727_1739, as well as the other related pathogenic proteins (GBAA1900, BA1900, BAS1762, BCE33L1711), are likely to be glycosyl hydrolase, BNR repeat-containing proteins that operate in the carbohydrate transport system PTS.

The pathogenic specific clusters A and L were the only other protein families that could be viewed in the context of their protein interactions. However, as none of the proteins in cluster A and L were found to be in a highly connected regions of the network, their interactions were considered in terms of their first-neighbours.

The species-specific interactions of the proteins in cluster A, with exception to BT9727_2561 which is not present in the *B. thuringiensis konkukian* (strain 97-27) PFIN, show a high level of similarity, illustrated by the vast number of orthologous connections. The level of similarity is greatest across the three *B. anthracis* subnetworks, with 22-23 proteins each. A significant proportion of these proteins are recognised as being involved in organic acid metabolism, including antibiotic responsive proteins. Interestingly, within the cross-species subnetwork, there exists four outsider proteins (GBAA3330, BA3330, BAS3087 and BCE33L2980); these have been annotated as penicillin acylase proteins. Links exist between the four unknown

pathogen specific proteins and these penicillin acylases through Interpro domain enrichment evidence. Earlier analysis of this protein family, detailed in section 4.2.3.3, also revealed the presence of a beta-lactamase-like region in all but BT9727_2561. In conjunction with the identification of penicillin acylase proteins it is likely the functions of these proteins relate to penicillin synthesis. The function of the remaining protein in this group, BT9727_2561, can potentially be inferred on the basis of these findings; although the consequence of not having a beta-lactamase-like region may alter the behaviour of this protein.

Sequence analysis of the pathogenic protein family L had uncovered no information about the probable functions of these putative extracellular proteins. However, PFIN analysis reveals they are likely to be involved in aromatic amino acid metabolism, based on their neighbouring proteins. Although showing this general behaviour, the first-neighbour networks of these proteins are quite dispersed across species, suggesting mechanisms of metabolism are quite different. It has previously been identified that the greatest difference between species in regards to aromatic amino acid metabolism is the mechanisms they use to regulate their expression [Sonenshein et al., 2002]. Based on this deduction, it could therefore be suggested that these pathogenic specific proteins are involved in this variable regulation mechanism.

The core lipoprotein family, PrsA, was also analysed, but in much greater depth. This major extracytoplasmic folding factor has been shown to effect the function of some essential translocated proteins involved in cell wall synthesis and functions relating to the cell membrane [Tjalsma et al., 2000, Wahlström et al., 2003]. Initial analysis of this protein family revealed the threshold used to construct the original PFINs was set too high, excluding a large proportion of functionally related proteins from the species-specific subnetworks, which in turn resulted in a minimal cross-species PrsA subnetwork.

Resetting the threshold to a weight of two created a more detailed view of the functional associations of PrsA across species. It can therefore be concluded that the original threshold of five, used to construct all the cross-species PFINs, removed a large proportion of the interactions. Analysis of the pathogenic and non-pathogenic

families using the newly refined species-specific PFINs could provide additional insight to that which has just been documented.

Also, as demonstrated by the PrsA cross-species PFIN, some expected orthologue links are not identified by simply inferring orthology based on the main orthologue pairs i.e. excluding paralogues. The inference of orthologues therefore requires an improved strategy in order to increase the level of detail captured through the cross-species PFINs generated.

Through analysis of the cross-species PrsA PFIN information concerning the interactions of PrsA could be determined. Despite much being known about the genomics of PrsA, the interactions in which it is involved are still not clear. However, in *B. subtilis*, it appears that PrsA is tightly coupled to other known heat shock stress response proteins (ClpP, ClpX, GroEL, GroES), as well as the functionally unknown protein YwrO. In addition, within this cross-species PFIN, a number of genes of unknown function interact with other known stress response proteins Ctc and Gsi.

Analysis of the *B. subtilis* cluster in the PrsA PFIN reveals the uncharacterised protein YacD is associated with many ribosomal proteins, stress response proteins, and proteins involved in protein folding. There are therefore potential targets identified in this PrsA PFIN for further directed studies, in particular with respect to *B. subtilis*.

An advantage of the approach that has been implemented is that it is easily extensible to incorporate new species and/or new datasets. Furthermore, alternative properties can be investigated. Here the use of PFINs to investigate the secretome was demonstrated, but other properties of interest can also be explored, whether that be the interactions of a single protein, or a larger collection of proteins.

Chapter 7

General discussion, conclusions and future work

This study has explored the application of e-science technology, in particular Web-service based workflows, to the analysis of microbial genomes using secretory protein prediction and subsequent protein family clustering. The comparison of clusters between organisms possessing different phenotypic properties (from pathogens to soil living bacteria) has revealed novel protein families that are potentially important in environmental adaptation. In particular a number of protein families were identified that could be potential virulent determinants for the pathogen *B. anthracis* and close relatives. Despite extensive conventional analysis, many protein families identified emerged from this analysis without any clues to their function. Methods for the integrative analysis of functional genomic data were then applied to develop functional integrated networks for all 11 *Bacillus* isolates whose complete genome sequence was available at the time of the beginning of the study. These networks were constructed for each species, and then integrated across species using orthology links, to facilitate the functional analysis of the protein families of unknown function. Finally, the value of cross-species probabilistic functional integrated networks for protein functional inference was demonstrated through their application to a case study of the protein family containing the secreted protein chaperone protein PrsA.

7.1 Discussion

The number of publicly available workflows has risen rapidly with the availability of accessible Web services and the development of Grid middleware and workflow enactors. Hundreds of workflows are available for use and new repositories are coming online to promote workflow sharing [Goble and Roure, 2007]. The BaSPP system presented here provides a novel set of workflows for the analysis of secreted proteins and provides yet another example of the value of an e-science approach to bioinformatics and genomic analysis. The ability to rapidly piece together annotation and protein analysis pipelines in a much more rapid and flexible fashion than hand coded applications is now becoming an accepted feature of e-science workflows [Alpdemir et al., 2005, Li et al., 2004, Stevens et al., 2004a,b].

However, in addition to providing a useful tool for biologists, the BaSPP system has also played a role in highlighting some of the areas of e-science that require further research. Firstly, the available workflow enactors do not provide interfaces that are flexible enough to customise for a particular application for a biologist in terms of the data querying and result display interface. Hence, it was necessary to design a Web portal that would allow a biologist to review and annotate the results of running the prediction and analysis workflows.

Secondly, BaSPP attempts to deal with the three key e-science problems: distribution, autonomy and heterogeneity. There were limited problems associated with distribution, the main difficulty arising as a result of the relatively large datasets that were being handled.

However, in many cases licensing issues will prevent a distributed approach from being adopted and complicate the issue of workflow sharing. Many legacy applications with licensing restrictions cannot be exposed to the outside world as services. Whilst Grid technology has promised to address such issues, practical systems to validate the rights of a user to access a service have yet to emerge. Service reliability was also a persistent problem in this study. As a result of these factors, many of the required services were implemented in-house, although still delivered using the Web services

infrastructure.

Dealing with data heterogeneity was less of a problem for this study. The services in the workflows mostly exchanged FASTA formatted protein lists. Using a custom database and parsing code removed any other issues related with heterogeneity within the application. In particular, the Grid based system, Microbase, was used to provide previously processed result sets. Using Microbase, in which comparison of the bacterial proteins has already been done and stored in a database, rapidly speeded up the operation of the analysis workflow, providing an insight into what can be achieved through interoperability of Grid based systems.

The putative *Bacillus* secretome data generated by the BaSPP workflows is also of great interest to microbiologists. The results of categorisation show a high degree of correlation to comparable bioinformatical analyses and validated experimental data. The analysis of the secreted protein families also provided some interesting results, particularly with respect to the pathogen-specific families. Laboratory verification of these findings is needed using mutant, overexpression and microarray experiments.

However, a large proportion of the secretome is of unknown function. In the *Bacillus* isolates categorised by BaSPP, there were 297 putative secreted families which were classified as miscellaneous. This work therefore provides data to prime further, biologically- and computationally-oriented investigations, to investigate the properties of these uncharacterised proteins, but also to confirm the predicted functions uncovered from the analysis.

The BaSPP package can be easily applied to the analysis of other Gram-positive bacteria, as demonstrated by application to *Staphylococcus aureus*, and with minor tweaking can even be applied to Gram-negative bacteria. All newly analysed data can be viewed and curated via the Web interface.

The concept of combining current bioinformatics programs is not a novel idea. The classification workflow of BaSPP is generally a sequential process, in which the predictions of different types of proteins occur at distinct steps in the workflow. However, three transmembrane tools are combined in parallel execution, which are subsequently integrated in order to make a prediction based on whether one or more

programs make a positive transmembrane prediction. Due to the resulting inaccuracy from incorporating MEMSAT and TMAP, the final decision rests solely with TMHMM. However, the concept of combining the predictions of different tools to make an overall decision was demonstrated. The strategy employed here resembles Phobius¹, in which transmembrane protein topology and signal peptide prediction is combined using TMHMM and SignalP respectively. Phobius demonstrates a more sophisticated method, using an HMM, to integrate the results from multiple tools to make an overall prediction on the secretory nature of the protein [Käll et al., 2007].

Secretory protein databases that use standard hand coded applications have also been reported previously. Secreted Protein Database² (SPD) is a collection of secreted proteins from Human, Mouse and Rat proteomes. As well as manually extracting secreted proteins from SWISSPROT based on their subcellular information, a bioinformatics workflow was also used, which incorporates the programs PSORT [Nakai and Horton, 1999], results of which are filtered using TMHMM, and Sec-HMMER (which uses a combination of CJ-SPHMM [Chen et al., 2003] and TMHMM to predict signal peptides and transmembrane regions respectively) [Chen et al., 2005].

In addition, a system called Augur has been developed by Billion and co-workers which also automates the process of categorising microbial surface proteins using a workflow, displaying the results via a Web interface [Billion et al., 2006]. The workflow is made up of four modules: orthologue detection (BLASTp), surface protein prediction (SignalP, TMHMM, HMMs from Superfamily [Gough et al., 2001] and Pfam [Finn et al., 2006] for cell surface motifs, and pattern matching for lipoproteins), COG classification to assign genes to functional groups based on BLAST, and SCOP classification [Murzin et al., 1995]. The results are subsequently stored in a database.

Unlike Augur, which identifies cell wall binding repeats and covalent attachment to the cell wall (LPXTG motif), BaSPP focuses solely on the covalent attachment of surface proteins. Covalently linked surface proteins play a key role in the display of surface proteins. The majority of these covalently attached cell wall binding pro-

¹Phobius Website: <http://phobius.cgb.ki.se>

²SPD Website: <http://spd.cbi.pku.edu.cn/>

teins are characterised by the sorting signal LPXTG [Cossart and Jonquières, 2000]. However, fewer LPXTG containing proteins were identified than had been anticipated. This may be accounted for by the identification of an additional C-terminal cell wall sorting signal NPQTN in *S. aureus*, *B. halodurans*, and *B. anthracis*, structurally related to the LPXTG motif [Boekhorst et al., 2005, Tjalsma et al., 2004]. Other identified sortase substrates include an NAKTN motif [Bierne et al., 2004] and QVPTGV motif [Barnett et al., 2004]. Therefore, searching for this motif would most likely have increased the count of the cell wall covalently attached proteins.

Augur and SPD are similar in operation to BaSPP, the significant difference being that BaSPP uses a service-oriented approach via which users with no programming knowledge can execute the workflow over their own datasets, as well as allowing the user to view and curate the results via a Web browser; Augur and SPD only provide a means of viewing already analysed data. BaSPP is also much more flexible. Services can be re-ordered, added and removed much more easily than a hard coded workflow, and remote programs and services incorporated more easily.

PFINs have proved to be very useful for predicting protein function in many different organisms including *S. cerevisiae* [Deng et al., 2004, Lee et al., 2004, Myers et al., 2005, Stark et al., 2006], Human [Rhodes et al., 2005, Xia et al., 2006] and others [Date and Stoeckert, 2006, von Mering et al., 2007]. To date the *Bacillus* community have not had the benefit of such a resource. In this study PFINs have been constructed for 11 different organisms from the genus *Bacillus*. The PFINs constructed for the *Bacillus* species provides an insight into biological mechanisms that would otherwise not be so easily identified/predicted. Specifically, data concerning unknown and previously unidentified links in the PFINs can be determined, from which functional predictions and relations can be judged. Functional inference can be achieved through viewing and querying the PFINs individually, but also via cross-species comparison. The development, analysis and comparison of the *Bacillus* PFINs extend current studies which have largely focused on eukaryotes. Furthermore, the scale of analysis has been increased to extend across a genus.

The integration of a variety of data sources to build the PFINs was important in

ensuring a large degree of coverage in terms of the size of the proteome, as the coverage of each experimental dataset linking pairs of genes differs. Each experiment has different associated error rates too [Lee et al., 2004]. Therefore applying probabilistic methods of data integration strengthens links that are common and weakens those that are not. The data sources available for *Bacillus* vary in their quality and size. In addition, there have been far fewer genomic scale studies on *Bacillus* species than for organisms such as *S. cerevisiae*. The PFINs constructed for *Bacillus* do not incorporate physical protein interaction data since no publicly available high-throughput experiments have been carried out. In contrast to other model organisms (e.g. *S. cerevisiae*), there are fewer high-throughput omic screens for the *Bacillus* isolates.

The use of a gold standard to uniformly weight the experimental data is not ideal. This approach not only prevents the inclusion of the gold standard dataset into the network, but the final weightings associated with each link between a pair of genes is dependent on the accuracy and coverage of this benchmark. Therefore, better, more sophisticated approaches are needed to weight experimental datasets for integration.

Comparison of the global structure of the PFINs visually showed the commonality in their structures, especially amongst the closely related isolates. Analysis of the network topology of the PFINs revealed the *Bacillus* PFINs to be small-world, scale-free, modular networks.

Using the *Bacillus* PFINs developed, the properties of some interesting secreted protein families were investigated, including those specific to the pathogens and non-pathogens, as well as the core family PrsA, by developing cross-species PFINs. It is currently possible to integrate across species based on the individual protein families identified by BaSPP, using clusters or first-neighbour based networks. Due to the size and complexity of the *Bacillus* PFINs, a global network alignment is difficult. However, newer network algorithms are now appearing that may make this possible [Cootes et al., 2007, Flannick et al., 2006, Hirsh and Sharan, 2007, Koyutürk et al., 2004, Sharan et al., 2005].

The database contains the data and metadata used in constructing each *Bacillus* PFIN. Through the course of development, the schema underwent continued changes

as new constraints were placed on it. The resulting schema is to form the core of the continued development in this area, with the planned construction of yeast and human PFINs.

The use of PFINs to complement sequence homology to infer protein function is now being recognised. Methods to provide this capability over PFINs are being developed. An example is documented in Chua et al. [2007], describing the inference of predicted Gene Ontology functional annotations using indirect networks. As modelling and analytical techniques continue to advance, these PFINs can be exploited further to explore biological mechanisms. The set of PFINs developed here promise to be of great value to the *Bacillus* community.

Throughout this work, whether that be in regards to BaSPP or SubtilNet, the focus has been on sharing and integrating data, collaborating between bioinformatician and biologists, in order to optimise the systems developed, and to be able to extract meaning from the results. To facilitate and help expand this process, e-science will be valuable. In the omic era, with the ever increasing amount of data available for analysis, e-science techniques will provide benefits in both dealing with scale and promoting collaboration required for improving the functional annotation of proteins.

7.2 Conclusions

In conclusion, the investigation of predicted bacterial secretomes through the use of workflows and e-science technology, as well as within and between species using PFIN modelling techniques, indicates the potential use of computing resources for maximising the information gained as multiple genomic sequences are generated through data integration. The BaSPP and SubtilNet frameworks developed provide a foundation on which other studies can be conducted, for a range of organisms. The knowledge gained from large scale analysis of secretomes can be used to generate inferences about bacterial evolution and niche adaptation. Modelling biological systems using PFINs provides a means of expanding component analysis to the broader context of the system, as demonstrated within this thesis through application to the secre-

tome. Both systems, whether used individually or in combination, can potentially lead to hypotheses to inform future experimental studies, that can ultimately result in a greater understanding of biology.

7.3 Future Work

One major future enhancement to this work relates to the e-science workflows developed. By using a notification service in conjunction with the workflows would provide a way of automatically analysing recently annotated genomes for secretory proteins. The need for users to interact with the workflow would therefore be removed, providing automatic updates of the database with new secretory protein data. Microbase could be used for this purpose, which essentially provides the necessary framework, enabling secretory protein categorisation and analysis to be added as a plugin.

In addition, the ability to migrate workflows and services closer to the data would provide a significant advantage; often the service executables are smaller than the data they operate over. However, this would be restricted due to the licensing constraints of many of the tools used. Therefore, an alternative would be to investigate the migration of workflows from the machine of their conception, to a remote enclosed environment from which they are able to access services that are unable to be exposed directly to third parties, and again Microbase could be utilised. Overall, the benefits of developing a way of implementing a migration process would be most beneficial.

There were limited problems associated with distribution, the main difficulty arising as a result of the relatively large datasets that were being handled. Improved data transport facilities, enabling transfer without SOAP packaging, as well as direct transfer between third parties, would increase the speed of the workflow execution. The new Taverna2 architecture should enable these functionalities [Oinn and Pocock pers. comm].

The effect of using the latest versions of tools in the classification workflow could also be explored in relation to the final putative secretomes. New improved versions SignalP 3.0 [Bendtsen et al., 2004] and MEMSAT3 [Jones, 2007] could be added.

A further enhancement would be to extend the classification process to identify cell wall non-covalently attached proteins. These proteins often bind via a number of repeated domains, including choline-binding domain, NlpC/P60 domain, LRR (leucine rich repeat), LysM domain, GW modules, S-layer homology motif [Desvaux et al., 2006]. Currently, searching for these cell wall binding repeats is outside the scope of BaSPP, but providing this functionality would be an obvious extension. It may also be interesting to add pattern recognition for the cell wall covalent attachment of proteins containing a NPQTN motif, previously identified in *S. aureus*, *B. halodurans*, and *B. anthracis* [Boekhorst et al., 2005], as well as perhaps the NAKTN motif and QVPTGV motifs.

In tandem with many other ongoing systems biology projects, enhancements can also be applied to the development and analysis of PFINs. An obvious addition would be to use more data sources to generate the networks. New data sources for inclusion may be new external datasets not previously incorporated; motifs identified upstream of genes would be one such dataset. Identification of these motifs in-house by simply analysing the genome means this source of data is applicable to any organism. Alternatively, data generated through analysing the PFIN itself may be plugged back into the network, reinforcing links thought to be important; the identification of context links would be one such case worth investigating, as utilised in Lee et al. [2004].

In building PFINs, the gold standard plays a central role. An investigation could therefore be undertaken into the development of a more substantial gold standard dataset with a wider coverage of protein-protein interactions. However, as the use of a gold standard removes the incorporation of this dataset from integration into the PFIN, a better approach would be to weight the data without the need for a gold standard.

The predictive capability and biases of the experimental datasets in relation to their contribution to the final network could also be assessed. The contribution of each experimental dataset individually to the network, as well the contribution to the network of different combinations of datasets, could be analysed, potentially revealing some other properties of the data being used.

Recently, the analysis of PFINs across species is also an area of great interest. The utilisation of the clustering algorithm MCODE to identify regions of high connectivity could be modified in order to use the alternative, and best clustering algorithm (according to Brohée and van Helden [2006]), MCL. In addition the application of alternative methods previously discussed in section 2.4.2 could also be utilised and potentially built into the framework.

It would also be interesting in the future to extend the workflow concept to the construction of PFINs. Due to the computational intensity of evaluating large datasets during the development process, the lack of distribution was an obvious benefit. Using a single host environment for execution and storage reduced the drawbacks associated with moving large datasets. However, using a single host relies largely on the availability of a platform capable of hosting this environment. Furthermore, enhancements may also be needed to improve performance depending on the growth of the infrastructure. Currently, many of the tasks executed in regards to PFIN construction and comparison were very time-consuming. For example, the integration of the different data sources for a particular organism takes approximately one day. The development of e-science platforms for PFIN development and comparison could therefore be a valuable contribution to this field, in which Grid resources could also be utilised. A user could therefore utilise Grid technology to remotely add new sources, integrate all or selected sources, to construct the PFIN, and finally view and query the PFIN. For example, the Grid resource Microbase was a valuable commodity in developing SubtilNet, in which pre-computed BLAST results were simply extracted, avoiding the need to do time-consuming BLAST searches.

An additional Cytoscape plugin could be developed to allow SubtilNet data to be exported to Cytoscape file formats at the click of a button. This could include single PFINs, perhaps pre-computed and stored in the database, or cross-species subnetworks generated on-the-fly. Exporting Cytoscape files would vastly improve the usability of the SubtilNet framework, and make it more attractive to biologists.

The methodology developed for the *Bacillus* genomes is planned to be applied to other organisms, including yeast and human. Extension of the in-house data

warehouse to more species will allow other interesting comparisons to be made.

Therefore, there are many avenues for future work from the foundation systems provided here. The SubtilNet framework provides the basis for the most significant computational improvements and extensions, although additions to the BaSPP system could also be applied. In biological terms, BaSPP enables the identification of secreted proteins, such as those specific to the pathogens, that can be investigated further in the wet-lab. SubtilNet, on-the-other-hand, is much broader in application, providing an extensive resource for biologists and bioinformaticians to investigate many different biological phenomena, which again, can be tested in the laboratory.

Bibliography

- MCODE documentation. URL <http://baderlab.org/Software/MCODE/UsersManual#head-d622ee68bd32104daaaed346dae4ed596728c31c>. Accessed 16 Jul 2007.
- Understanding Metadata*. NISO Press, 2004. URL www.niso.org. Accessed 8 Sept 07.
- BioDAS. Accessed 3 Sept 07. URL http://www.biodas.org/wiki/Main_Page.
- M. Addis, J. Ferris, M. Greenwood, P. Li, D. Marvin, T. Oinn, and A. Wipat. Experiences with e-Science workflow specification and enactment in bioinformatics. In *OST e-Science Second All Hands Meeting (AHM'09)*, 2003.
- A. Aderem. Systems biology: its practice and challenges. *Cell*, 121(4):511–513, May 2005.
- A. Alexeyenko, I. Tamas, G. Liu, and E. L. L. Sonnhammer. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22(14): e9–15, Jul 2006.
- N. E. E. Allenby, N. O'Connor, Z. Prgai, N. M. Carter, M. Miethke, S. Engelmann, M. Hecker, A. Wipat, A. C. Ward, and C. R. Harwood. Post-transcriptional regulation of the *Bacillus subtilis* pst operon encoding a phosphate-specific ABC transporter. *Microbiology*, 150(Pt 8):2619–2628, Aug 2004.
- M. N. Alpdemir, A. Mukherjee, N. W. Paton, A. A.A. Fernandes, P. Watson, K. Glover, C. Greenhalgh, T. Oinn, and H. Tipney. *Advances in Grid Computing - EGC 2005 Monday, July 11,*, volume Volume 3470/2005 of *Lecture Notes in*

- Computer Science*, chapter Contextualised Workflow Execution in MyGrid, pages 444–453. Springer Berlin / Heidelberg, 2005.
- S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.
- M. Altunay, D. Colonnese, and C. Warade. Web services for bioinformatics. May 2004. URL <http://www-128.ibm.com/developerworks/webservices/library/ws-bioinfo.html>. Accessed 30 Apr 2007.
- G. L. Andersen, J. M. Simchock, and K. H. Wilson. Identification of a region of genetic variability among *Bacillus anthracis* strains and related species. *J Bacteriol*, 178(2):377–384, Jan 1996.
- T. Andrews, F. Curbera, H. Dholakia, Y. Goland, J. Klein, F. Leymann, K. Liu, D. Roller, D. Smith, S. Thatte, I. Trickovic, and S. Weerawarana. Business Process Execution Language for Web Services Version 1.1. May 2003. URL <http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-bpel/ws-bpel.pdf>. Accessed 18 Jun 2007.
- G. M. Anlezark, T. Vaughan, E. Fashola-Stone, N. P. Michael, H. Murdoch, M. A. Sims, S. Stubbs, S. Wigley, and N. P. Minton. *Bacillus amyloliquefaciens* orthologue of *bacillus subtilis* ywro encodes a nitroreductase enzyme which activates the prodrug cb 1954. *Microbiology*, 148(Pt 1):297–306, Jan 2002.
- H. Antelmann, C. Scharf, and M. Hecker. Phosphate starvation-inducible proteins of *Bacillus subtilis*: proteomics and transcriptional analysis. *J Bacteriol*, 182(16):4478–4490, Aug 2000.
- H. Antelmann, H. Tjalsma, B. Voigt, S. Ohlmeier, S. Bron, J. M. van Dijl, and M. Hecker. A proteomic view on genome-based signal peptide predictions. *Genome Res.*, 11(9):1484–502, Sep 2001.
- L. Aravind. Guilt by association: contextual information in genome analysis. *Genome Res*, 10(8):1074–1077, Aug 2000.

- J. S. V. Arnam, J. L. McMurry, M. Kihara, and R. M. Macnab. Analysis of an engineered salmonella flagellar fusion protein, flir-flhb. *J Bacteriol*, 186(8):2495–2498, Apr 2004.
- Y. Assenov. Topological analysis of biological networks. Master’s thesis, Universität des Saarlandes, November 2006.
- S. Asur, D. Ucar, and S. Parthasarathy. An ensemble framework for clustering protein-protein interaction networks. *Bioinformatics*, 23(13):i29–i40, Jul 2007.
- T. K. Attwood and D. J. Parry-Smith. *Introduction to Bioinformatics*. Prentice Hall, 1999.
- G. D. Bader and C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, Jan 2003.
- S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Res*, 16(3):428–435, Mar 2006.
- A. L. Barabási and Z. N. Oltvai. Network biology: understanding the cell’s functional organization. *Nat Rev Genet*, 5(2):101–13, Feb 2004.
- T. C. Barnett, A. R. Patel, and J. R. Scott. A novel sortase, srtc2, from streptococcus pyogenes anchors a surface protein containing a qvptgv motif to the cell wall. *J Bacteriol*, 186(17):5865–5875, Sep 2004.
- V. V. Bartsevich and H. B. Pakrasi. Molecular identification of an ABC transporter complex for manganese: analysis of a cyanobacterial mutant strain impaired in the photosynthetic oxygen evolution process. *EMBO J*, 14(9):1845–1853, May 1995.
- V. Batagelj and A. Mrvar. Pajek – program for large network analysis. *Connections*, 21:47–57, 1998.

- J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–795, Jul 2004.
- J. D. Bendtsen, H. Nielsen, D. Widdick, T. Palmer, and S. Brunak. Prediction of twin-arginine signal peptides. *BMC Bioinformatics*, 6:167, 2005.
- J. Berg and M. Lässig. Local graph alignment and motif search in biological networks. *Proc Natl Acad Sci U S A*, 101(41):14689–14694, Oct 2004.
- B. C. Berks, T. Palmer, and F. Sargent. Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. *Curr Opin Microbiol*, 8(2):174–181, Apr 2005.
- E. Biemans-Oldehinkel, M. K. Doeven, and B. Poolman. ABC transporter architecture and regulatory roles of accessory domains. *FEBS Lett*, 580(4):1023–1035, Feb 2006.
- H. Bierne, C. Garandeau, M. G. Pucciarelli, C. Sabet, S. Newton, F. Garcia del Portillo, P. Cossart, and A. Charbit. Sortase b, a new class of sortase in *listeria monocytogenes*. *J Bacteriol*, 186(7):1972–1982, Apr 2004.
- A. Billion, R. Ghai, T. Chakraborty, and T. Hain. Augur—a computational pipeline for whole genome microbial surface protein prediction and classification. *Bioinformatics*, 22(22):2819–2820, Nov 2006.
- T. T. Binnewies, J. D. Bendtsen, P. F. Hallin, N. Nielsen, T. M. Wassenaar, M. B. Pedersen, P. Klemm, and D. W. Ussery. Genome update: Protein secretion systems in 225 bacterial genomes. *Microbiology*, 151(Pt 4):1013–1016, Apr 2005.
- M. Blatt, S. Wiseman, and E. Domany. Superparamagnetic clustering of data. *Phys Rev Lett*, 76(18):3251–3254, Apr 1996.
- J. D. Blower, A. B. Harrison, and K. Haines. *Computational Science ICCS 2006*, volume 3993/2006 of *Lecture Notes in Computer Science*, chapter Styx Grid Ser-

- vices: *Lightweight, Easy-to-Use Middleware for Scientific Workflows*, pages 996–1003. Springer Berlin / Heidelberg, 2006.
- J. Boekhorst, M. W. H. J. de Been, M. Kleerebezem, and R. J. Siezen. Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs. *J Bacteriol*, 187(14):4928–4934, Jul 2005.
- J. Boekhorst, M. Wels, M. Kleerebezem, and R. J. Siezen. The predicted secretome of *Lactobacillus plantarum* WCFS1 sheds light on interactions with its environment. *Microbiology*, 152(Pt 11):3175–3183, Nov 2006.
- A. Bolhuis, C. P. Broekhuizen, A. Sorokin, M. L. van Roosmalen, G. Venema, S. Bron, W. J. Quax, and J. M. van Dijl. SecDF of *Bacillus subtilis*, a molecular Siamese twin required for the efficient secretion of proteins. *J Biol Chem*, 273(33):21217–21224, Aug 1998.
- A. Bolhuis, A. Matzen, H. L. Hyyrylinen, V. P. Kontinen, R. Meima, J. Chapuis, G. Venema, S. Bron, R. Freudl, and J. M. van Dijl. Signal peptide peptidase and clpp-like proteins of *Bacillus subtilis* required for efficient translocation and processing of secretory proteins. *J Biol Chem*, 274(35):24585–24592, Aug 1999.
- D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, C. Ferris, and D. Orchard. Web Service Architecture. Technical report, World Wide Web Consortium (W3C), 2004. URL www.w3.org/TR/ws-arch/. Accessed 18 Jun 2007.
- A. Borges, C. F. Hawkins, L. C. Packman, and R. N. Perham. Cloning and sequence analysis of the genes encoding the dihydrolipoamide acetyltransferase and dihydrolipoamide dehydrogenase components of the pyruvate dehydrogenase multienzyme complex of *Bacillus stearothermophilus*. *Eur J Biochem*, 194(1):95–102, Nov 1990.
- B. Breitkreutz, C. Stark, and M. Tyers. Osprey: a network visualization system. *Genome Biol*, 4(3):R22, 2003.

- P. Brodin, I. Rosenkrands, P. Andersen, S. T. Cole, and R. Brosch. ESAT-6 proteins: protective antigens and virulence factors? *Trends Microbiol*, 12(11):500–508, Nov 2004.
- S. Brohée and J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488, 2006.
- K. R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, May 2005.
- G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili. Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, 433(7025):531–537, Feb 2005.
- S. Calogero, R. Gardan, P. Glaser, J. Schweizer, G. Rapoport, and M. Debarbouille. RocR, a novel regulatory protein controlling arginine utilization in *Bacillus subtilis*, belongs to the NtrC/NifA family of transcriptional activators. *J Bacteriol*, 176(5):1234–1241, Mar 1994.
- M. Campillos, C. von Mering, L. J. Jensen, and P. Bork. Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Res*, 16(3):374–382, Mar 2006. URL <http://dx.doi.org/10.1101/gr.4336406>.
- N. Campo, H. Tjalsma, G. Buist, D. Stepniak, M. Meijer, M. Veenhuis, M. Westermann, J. P. Mller, S. Bron, J. K., O. P. Kuipers, and J. D. H. Jongbloed. Subcellular sites for bacterial protein export. *Mol Microbiol*, 53(6):1583–1599, Sep 2004.
- G. Cenci, F. Trotta, and G. Caldini. Tolerance to challenges miming gastrointestinal transit by spores and vegetative cells of *Bacillus clausii*. *J Appl Microbiol*, 101(6):1208–1215, Dec 2006.
- CERN. GridCafé. URL <http://gridcafe.web.cern.ch/>. Accessed 30 Nov 2006.

- F. Chen, A. J. Mackey, J. K. Vermunt, and D. S. Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2:e383, 2007.
- I. Chen and D. Dubnau. DNA uptake during bacterial transformation. *Nat Rev Microbiol*, 2(3):241–249, Mar 2004.
- Y. Chen, P. Yu, J. Luo, and Y. Jiang. Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. *Mamm Genome*, 14(12):859–865, Dec 2003.
- Y. Chen, Y. Zhang, Y. Yin, G. Gao, S. Li, Y. Jiang, X. Gu, and J. Luo. Spd—a web-based secreted protein database. *Nucleic Acids Res*, 33(Database issue): D169–D173, Jan 2005.
- H. N. Chua, W. Sung, and L. Wong. Using indirect protein interactions for the prediction of gene ontology functions. *BMC Bioinformatics*, 8 Suppl 4:S8, 2007.
- D. Clery. Infrastructure. Can grid computing help us work together? *Science*, 313 (5786):433–434, Jul 2006.
- J. R. Cole, B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*, 33(Database issue):D294–D296, Jan 2005.
- A. P. Cootes, S. H. Muggleton, and M. J. E. Sternberg. The identification of similarities between biological networks: Application to the metabolome and interactome. *J Mol Biol*, 369(4):1126–1139, Jun 2007.
- R. R. Copley, R. B. Russell, and C. P. Ponting. Sialidase-like Asp-boxes: sequence-similar structures within different protein folds. *Protein Sci*, 10(2):285–292, Feb 2001.
- P. Cossart and R. Jonquières. Sortase, a universal target for therapeutic agents

- against Gram-positive bacteria? *Proc Natl Acad Sci U S A*, 97(10):5013–5015, May 2000.
- T. Craddock, P. Lord, C. Harwood, and A. Wipat. e-Science tools for the genomic scale characterisation of bacterial secreted proteins. In S. J. Cox, editor, *Proceedings of the UK e-Science All Hands Meeting 2006*, pages 788–795. National e-Science Centre, 2006.
- F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, and S. Weerawarana. Unraveling the Web services web: an introduction to SOAP, WSDL, and UDDI. *Internet Computing, IEEE*, 6(2):86–93, Mar/Apr 2002.
- V. Curwen, E. Eyraas, T. D. Andrews, L. Clarke, E. Mongin, S. M. J. Searle, and M. Clamp. The Ensembl automatic gene annotation system. *Genome Res*, 14(5): 942–950, May 2004.
- M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill. Interactome: gateway into systems biology. *Hum Mol Genet*, 14 Spec No. 2:R171–R181, Oct 2005.
- S. V. Date and C. J. Stoeckert. Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. *Genome Res*, 16(4):542–549, Apr 2006.
- M. W. Davidson and The Florida State University. Molecular expressions cell biology: Bacteria cell structure. Jan 2005. URL <http://micro.magnet.fsu.edu/cells/bacteriacell.html>. Accessed 21 Feb 2007.
- A. S. de Boer, F. Priest, and B. Diderichsen. On the industrial use of *Bacillus licheniformis*: a review. *Appl Microbiol Biotechnol*, 40(5):595–598, Jan 1994.
- E. de Leeuw, B. Graham, G. J. Phillips, C. M. ten Hagen-Jongman, B. Oudega, and J. Luirink. Molecular characterization of *Escherichia coli* FtsE and FtsX. *Mol Microbiol*, 31(3):983–993, Feb 1999.

- T. F. Deluca, I. Wu, J. Pu, T. Monaghan, L. Peshkin, S. Singh, and D. P. Wall. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, 22(16):2044–2046, Aug 2006.
- M. Deng, T. Chen, and F. Sun. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol*, 11(2-3):463–475, 2004.
- M. Desvaux, E. Dumas, I. Chafsey, and M. I. Hébraud. Protein cell surface display in gram-positive bacteria: from single protein to macromolecular protein structure. *FEMS Microbiol Lett*, 256(1):1–15, Mar 2006.
- E. Deuerling, A. Mogk, C. Richter, M. Purucker, and W. Schumann. The *ftsH* gene of *Bacillus subtilis* is involved in major cellular processes such as sporulation, stress adaptation and secretion. *Mol Microbiol*, 23(5):921–933, Mar 1997.
- T. C. Dixon, M. Meselson, J. Guillemin, and P. C. Hanna. Anthrax. *N Engl J Med*, 341(11):815–826, Sep 1999.
- A. Dove. Proteomics: translating genomics into products? *Nat Biotechnol*, 17(3):233–236, Mar 1999.
- S. Eder, L. Shi, K. Jensen, K. Yamane, and F. M. Hulett. A *Bacillus subtilis* secreted phosphodiesterase/alkaline phosphatase is the product of a *Pho* regulon gene, *phoD*. *Microbiology*, 142 (Pt 8):2041–2047, Aug 1996.
- M. Ehling-Schulz, M. Guinebretiere, A. Monthán, O. Berge, M. Fricker, and B. Svensson. Toxin gene profiling of enterotoxic and emetic *Bacillus cereus*. *FEMS Microbiol Lett*, 260(2):232–240, Jul 2006.
- A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575–1584, Apr 2002.
- A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, Nov 1999.

- A. J. Enright, V. Kunin, and C. A. Ouzounis. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res*, 31(15):4632–4638, Aug 2003.
- R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, Information and Computer Science, 2000.
- R. T. Fielding and R. N. Taylor. Principled design of the modern Web architecture. *ACM Trans. Inter. Tech.*, 2(2):115–150, 2002.
- R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–D251, Jan 2006.
- J. Flannick, A. Novak, B. S. Srinivasan, H. H. McAdams, and S. Batzoglou. Græmlin: general and robust alignment of multiple large interaction networks. *Genome Res*, 16(9):1169–1181, Sep 2006.
- I. Foster. *Grid Computing: Making the Global Infrastructure a Reality*, chapter The Grid: A New Infrastructure for 21st Century Science. Wiley Series in Communications Networking & Distributed Systems. 2003.
- I. Foster, C. Kesselman, and Steven Tuecke. The anatomy of the Grid:enabling scalable virtual organisations. *International Journal of Supercomputer Applications*, 15(3), Oct 2001.
- T. Gabaldón and M. A. Huynen. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci*, 61(7-8):930–944, Apr 2004.
- C. S. Galloway, P. Wang, D. Winstanley, and I. M. Jones. Comparison of the bacterial Enhancin-like proteins from *Yersinia* and *Bacillus* spp. with a baculovirus Enhancin. *J Invertebr Pathol*, 90(2):134–137, Oct 2005.
- T. K. B. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. S. Mohan, S.I Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nan-

- dakumar, B. Shen, N. Deshpande, R. Nayak, M. Sarker, J. D. Boeke, G. Parmigiani, J. Schultz, J. S. Bader, and A. Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3): 285–293, Mar 2006.
- R. Gardan, G. Rapoport, and M. Débarbouillé. Expression of the rocDEF operon involved in arginine catabolism in *Bacillus subtilis*. *J Mol Biol*, 249(5):843–856, Jun 1995.
- A. Gattiker, E. Gasteiger, and A. Bairoch. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics*, 1(2):107–108, 2002.
- H. Ge, A. J. M. Walhout, and M. Vidal. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*, 19(10):551–560, Oct 2003.
- Gene Ontology Consortium. GO slim and subset guide. a. URL <http://www.geneontology.org/GO.slims.shtml>. Accessed 7th Jun 2007.
- Gene Ontology Consortium. An introduction to the Gene Ontology. b. URL <http://www.geneontology.org/GO.doc.shtml>. Accessed 19 Feb 2007.
- C. A. Goble and D. C. De Roure. myExperiment: social networking for workflow-using e-scientists. In *WORKS '07: Proceedings of the 2nd workshop on Workflows in support of large-scale science*, pages 1–2, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-715-5.
- K. Gottschalk, S. Graham, H. Kreger, and J. Snell. Introduction to Web services architecture. *IBM Systems Journal*, 41(2):170–177, 2002.
- J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J Mol Biol*, 313(4):903–919, Nov 2001.
- J. Gray, D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber.

- Scientific data management in the coming decade. *SIGMOD Rec.*, 34(4):34–41, 2005. ISSN 0163-5808.
- M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau, and T. Oinn. Provenance of e-Science experiments - experience from bioinformatics. In *Proceedings UK e-Science All Hands Meeting*, 2003.
- M. Hajaij-Ellouze, S. Fedhila, D. Lereclus, and C. Nielsen-LeRoux. The enhancin-like metalloprotease from the *Bacillus cereus* group is regulated by the pleiotropic transcriptional activator PlcR but is not essential for larvicidal activity. *FEMS Microbiol Lett*, 260(1):9–16, Jul 2006.
- W. G. Hale, J. P. Margham, and V. A. Saunders. *Collins Dictionary of Biology*. HarperCollins, 1995.
- J. Hallinan and A. Wipat. Clustering and cross-talk in a yeast functional interaction network. In *IEEE Symposium on computational Intelligence in Bioinformatics and Computational Biology*, 2006.
- C. S. Han, G. Xie, J. F. Challacombe, M. R. Altherr, S. S. Bhotika, N. Brown, D. Bruce, C. S. Campbell, M. L. Campbell, J. Chen, O. Chertkov, C. Cleland, M. Dimitrijevic, N. A. Doggett, J. J. Fawcett, T. Glavina, L. A. Goodwin, L. D. Green, K. K. Hill, P. Hitchcock, P. J. Jackson, P. Keim, A. R. Kewalramani, J. Longmire, S. Lucas, S. Malfatti, K. McMurphy, L. J. Meincke, M. Misra, B. L. Moseman, M. Mundt, A. C. Munk, R. T. Okinaka, B. Parson-Quintana, L. P. Reilly, P. Richardson, D. L. Robinson, E. Rubin, E. Saunders, R. Tapia, J. G. Tesmer, N. Thayer, L. S. Thompson, H. Tice, L. O. Ticknor, P. L. Wills, T. S. Brettin, and P. Gilna. Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. *J Bacteriol*, 188(9):3382–3390, May 2006.
- D. W. Hanlon and G. W. Ordal. Cloning and characterization of genes encoding

- methyI-accepting chemotaxis proteins in *Bacillus subtilis*. *J Biol Chem*, 269(19): 14038–14046, May 1994.
- C. Harwood pers. comm.
- Health Protection Agency. *Bacillus* spp. URL http://www.hpa.org.uk/infections/topics_az/bacillus/menu.htm. Accessed 29 Jan 2007.
- M. Hecker and U. Völker. Non-specific, general and multiple stress resistance of growth-restricted *bacillus subtilis* cells by the expression of the sigma^{ab} regulon. *Mol Microbiol*, 29(5):1129–1136, Sep 1998.
- D. Heckerman. A tutorial on learning with Bayesian networks. Technical report, Microsoft Research, Nov 1996.
- T. Hernandez and S. Kambhampati. Integration of biological sources: Current systems and challenges. *ACM SIGMOD Record*, 33(3):51–60, Sep 2004.
- T. Hey and A. Trefethen. e-Science and its implications. *Philos Transact A Math Phys Eng Sci*, 361(1809):1809–1825, Aug 2003.
- T. Hey and A. E. Trefethen. Cyberinfrastructure for e-Science. *Science*, 308(5723): 817–821, May 2005.
- I. Hirose, K. Sano, I. Shioda, M. Kumano, K. Nakamura, and K. Yamane. Proteome analysis of *Bacillus subtilis* extracellular proteins: a two-dimensional protein electrophoretic study. *Microbiology*, 146 (Pt 1):65–75, Jan 2000.
- L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, Dec 2002.
- E. Hirsh and R. Sharan. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*, 23(2):e170–e176, Jan 2007.

- M. Hoebeke, H. Chiapello, P. Noirot, and P. Bessi res. SPiD: a subtilis protein interaction database. *Bioinformatics*, 17(12):1209–1212, Dec 2001.
- M. J. L. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pac Symp Biocomput*, pages 276–287, 2004.
- Z. Hu, J. Mellor, J. Wu, M. Kanehisa, J. M. Stuart, and C. Delisi. Towards zoomable multidimensional maps of the cell. *Nat Biotechnol*, 25(5):547–554, May 2007.
- F. M. Hulett, E. E. Kim, C. Bookstein, N. V. Kapp, C. W. Edwards, and H. W. Wyckoff. *Bacillus subtilis* alkaline phosphatases iii and iv. Cloning, sequencing, and comparisons of deduced amino acid sequence with *Escherichia coli* alkaline phosphatase three-dimensional structure. *J Biol Chem*, 266(2):1077–1084, Jan 1991.
- D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34(Web Server issue):W729–W732, Jul 2006.
- J. M. In cio, C. Costa, and I. de S -Nogueira. Distinct molecular mechanisms involved in carbon catabolite repression of the arabinose regulon in *bacillus subtilis*. *Microbiology*, 149(Pt 9):2345–2355, Sep 2003.
- N. Ivanova, A. Sorokin, I. Anderson, N. Galleron, B. Candelon, V. Kapatral, A. Bhat-tacharyya, G. Reznik, N. Mikhailova, A. Lapidus, L. Chu, M. Mazur, E. Goltsman, Larsen N., M. D’Souza, T. Walunas, Y. Grechkin, G. Pusch, R. Haselkorn, M. Fon-stein, S. D. Ehrlich, R. Overbeek, and N. Kyrpides. Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature*, 423(6935):87–91, May 2003.
- Jamstec. August 1999. URL <http://www.jamstec.go.jp/jamstec-e/PR/9908/0810b.html>. Accessed 23 Mar 2007.
- R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian networks approach

- for predicting protein-protein interactions from genomic data. *Science*, 302(5644): 449–453, Oct 2003.
- G. B. Jensen, B. M. Hansen, J. Eilenberg, and J. Mahillon. The hidden lifestyles of *Bacillus cereus* and relatives. *Environ Microbiol*, 5(8):631–640, Aug 2003.
- D. T. Jones. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5):538–544, Mar 2007.
- D. T. Jones, W. R. Taylor, and J. M. Thornton. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33(10):3038–3049, Mar 1994.
- K. Joung and J. Côté. Evaluation of ribosomal RNA gene restriction patterns for the classification of *Bacillus* species and related genera. *J Appl Microbiol*, 92(1): 97–108, 2002.
- A. R. Joyce and B. Ø. Palsson. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7(3):198–210, Mar 2006.
- A. S. Juncker, H. Willenbrock, G. Von Heijne, S. Brunak, H. Nielsen, and A. Krogh. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*, 12(8):1652–1662, Aug 2003.
- L. Käll, A. Krogh, and E. L. L. Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction—the phobius web server. *Nucleic Acids Res*, May 2007.
- M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):D354–D357, Jan 2006.
- R. M. Kappes, B. Kempf, S. Kneip, J. Boch, J. Gade, J. Meier-Wagner, and E. Bremer. Two evolutionarily closely related ABC transporters mediate the uptake of

- choline for synthesis of the osmoprotectant glycine betaine in *Bacillus subtilis*. *Mol Microbiol*, 32(1):203–216, Apr 1999.
- K. A. Karasavvas, R. Baldock, and A. Burger. Bioinformatics integration and agent technology. *J Biomed Inform*, 37(3):205–219, Jun 2004.
- B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A*, 100(20):11394–11399, Sep 2003.
- B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker. Path-BLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, 32(Web Server issue):W83–W88, Jul 2004.
- Kenneth Todar University of Wisconsin-Madison Department of Bacteriology. The genus *Bacillus*. *Todar's Online Textbook of Bacteriology*, 2005. URL <http://textbookofbacteriology.net/Bacillus.html>. Accessed 30 Jun 07.
- L. Kiemer, S. Costa, M. Ueffing, and G. Cesareni. WI-PHI: A weighted yeast interactome enriched for direct physical interactions. *Proteomics*, 7(6):932–943, Mar 2007.
- A. D. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, Nov 2004.
- H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, Mar 2002.
- A. Koide and J. A. Hoch. Identification of a second oligopeptide transport system in *Bacillus subtilis* and determination of its role in sporulation. *Mol Microbiol*, 13(3):417–426, Aug 1994.
- V. P. Kontinen, P. Saris, and M. Sarvas. A gene (*prsA*) of *Bacillus subtilis* involved in a novel, late stage of protein export. *Mol Microbiol*, 5(5):1273–1283, May 1991.

- M. Koyutürk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, 20 Suppl 1:I200–I207, Aug 2004.
- A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting trans-membrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, 305(3):567–580, Jan 2001.
- F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessières, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, and A. Danchin. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, 390(6657):249–256, Nov 1997.
- V. Lazarevic and D. Karamata. The tagGH operon of *Bacillus subtilis* 168 encodes a two-component ABC transporter involved in the metabolism of two wall teichoic acids. *Mol Microbiol*, 16(2):345–355, Apr 1995.
- T. F. Leal and I. de Sá-Nogueira. Purification, characterization and functional analysis of an endo-arabinanase (AbnA) from *Bacillus subtilis*. *FEMS Microbiol Lett*, 241(1):41–48, Dec 2004.
- I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558, Nov 2004.
- S. Lewenza, J. L. Gardy, F. S. L. Brinkman, and R. E. W. Hancock. Genome-wide identification of *Pseudomonas aeruginosa* exported proteins using a consensus computational strategy combined with a laboratory-based PhoA fusion screen. *Genome Res*, 15(2):321–329, Feb 2005.
- M. A. Leyva-Vazquez and P. Setlow. Cloning and nucleotide sequences of the genes encoding triose phosphate isomerase, phosphoglycerate mutase, and enolase from *Bacillus subtilis*. *J Bacteriol*, 176(13):3903–3910, Jul 1994.

- D. Li, J. Li, S. Ouyang, J. Wang, S. Wu, P. Wan, Y. Zhu, X. Xu, and F. He. Protein interaction networks of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*: Large-scale organization and robustness. *Proteomics*, 6(2):456–461, Nov 2006.
- L. Li, C. J. Stoeckert, and D. S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–2189, Sep 2003.
- P. Li, K. Hayward, C. Jennings, K. Owen, T. Oinn, R. Stevens, S. Pearce, and A. Wipat. Association of variations in I kappa B-epsilon with Graves' disease using classical and myGrid methodologies. In S. J. Cox, editor, *Proceedings of the UK e-Science All Hands Meeting*, 2004.
- K. J. Linton. Structure and function of ABC transporters. *Physiology (Bethesda)*, 22:122–130, Apr 2007.
- P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, Jul 2003.
- S. Maere, K. Heymans, and M. Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, Aug 2005.
- Y. Makita, M. Nakao, N. Ogasawara, and K. Nakai. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res*, 32(Database issue):D75–D77, Jan 2004.
- J. A. Malek, J. M. Wierzbowski, W. Tao, S. A. Bosak, D. J. Saranga, L. Doucette-Stamm, D. R. Smith, P. J. McEwan, and K. J. McKernan. Protein interaction mapping on a functional shotgun sequence of *rickettsia sibirica*. *Nucleic Acid Res*, 32(3):1059–1064, Feb 2004.

- L. A. Marraffini, A. C. Dedent, and O. Schneewind. Sortases and the art of anchoring proteins to the envelopes of Gram-positive bacteria. *Microbiol Mol Biol Rev*, 70 (1):192–221, Mar 2006.
- C. Mathiopoulos, J. P. Mueller, F. J. Slack, C. G. Murphy, S. Patankar, G. Bukusoglu, and A. L. Sonenshein. A *Bacillus subtilis* dipeptide transport system expressed early during sporulation. *Mol Microbiol*, 5(8):1903–1913, Aug 1991.
- L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, 11(12):2120–2126, Dec 2001.
- J. A. McDonough, K. E. Hacker, A. R. Flores, M. S. Pavelka, and M. Braunstein. The twin-arginine translocation pathway of *Mycobacterium smegmatis* is functional and required for the export of mycobacterial beta-lactamases. *J Bacteriol*, 187(22): 7667–7679, Nov 2005.
- J. C. Mellor, I. Yanai, K. I. H. Clodfelter, J. Mintseris, and C. DeLisi. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*, 30(1): 306–309, Jan 2002.
- Microbase. URL <http://www.microbase.org.uk/>. Accessed 20 Sept 2007.
- N. Mitra and Y. Lafon. SOAP Version 1.2 Part 0: Primer (Second Edition). April 2007. URL <http://www.w3.org/TR/soap12-part0/>. Accessed 8 Jun 2007.
- S. Möller, M. D. Croning, and R. Apweiler. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17(7):646–653, Jul 2001.
- I. Moszer, L. M. Jones, S. Moreira, C. Fabry, and A. Danchin. SubtiList: the reference database for the *Bacillus subtilis* genome. *Nucleic Acids Res*, 30(1):62–65, Jan 2002.
- N. Murphy, D. J. McConnell, and B. A. Cantwell. The DNA sequence of the gene and

- genetic control sites for the excreted *B. subtilis* enzyme beta-glucanase. *Nucleic Acids Res*, 12(13):5355–5367, Jul 1984.
- A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995.
- C. L. Myers, D. Robson, A. Wible, M. A. Hibbs, C. Chiriac, C. L. Theesfeld, K. Dolinski, and O. G. Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome Biol*, 6(13):R114, 2005.
- K. Nakai and P. Horton. Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24(1):34–36, Jan 1999.
- W. W. Navarre and O. Schneewind. Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol Mol Biol Rev*, 63(1):174–229, Mar 1999.
- NCBI Entrez Genome Project Database. URL <http://www.ncbi.nih.gov/sites/entrez?db=genomeprj>. Accessed 29 Jun 2007.
- H. Nielsen and A. Krogh. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol. (ISMB 6)*, 6:122–130, 1998.
- P. Noirot and M. Noirot-Gros. Protein interaction networks in bacteria. *Curr Opin Microbiol*, 7(5):505–512, Oct 2004.
- M. Noirot-Gros, E. Dervyn, L. J. Wu, P. Mervelet, J. Errington, S. D. Ehrlich, and P. Noirot. An expanded view of bacterial dna replication. *Proc Natl Acad Sci U S A*, 99(12):8342–8347, Jun 2002.
- K. Nomura and S. Y. He. Powerful screens for bacterial virulence proteins. *Proc Natl Acad Sci U S A*, 102(10):3527–3528, Mar 2005.
- P. Nurse. Systems biology: understanding cells. *Nature*, 424(6951):883, Aug 2003.

- K. P. O'Brien, M. Remm, and E. L. L. Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):D476–D480, Jan 2005.
- H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*, 28(20):4021–4028, Oct 2000.
- T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, Nov 2004.
- T. Oinn, M. Greenwood, M. Addis, M. N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. R. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe. Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 2006.
- T. Oinn and M. Pocock pers. comm.
- S. Okuda, T. Katayama, S. Kawashima, S. Goto, and M. Kanehisa. ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res*, 34(Database issue):D358–D362, Jan 2006.
- S. Oliver. Proteomics: Guilt-by-association goes global. *Nature*, 403(6770):601–603, Feb 2000.
- G. J. Olsen and C. R. Woese. Ribosomal RNA: a key to phylogeny. *FASEB J*, 7(1):113–123, Jan 1993.
- R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 96(6):2896–2901, Mar 1999.

- M. J. Pallen. The ESAT-6/WXG100 superfamily – and a new Gram-positive secretion system? *Trends Microbiol*, 10(5):209–212, May 2002.
- Y. Park and C. Park. Topology of RbsC, a membrane component of the ribose transporter, belonging to the AraH superfamily. *J Bacteriol*, 181(3):1039–1042, Feb 1999.
- J. Parkhill, B. W. Wren, N. R. Thomson, R. W. Titball, M. T. Holden, M. B. Prentice, M. Sebahia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, A. M. Cerde no Tárraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Leather, S. Moule, P. C. Oyston, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, 413(6855):523–527, Oct 2001.
- M. Perego, P. Glaser, A. Minutello, M. A. Strauch, K. Leopold, and W. Fischer. Incorporation of D-alanine into lipoteichoic acid and wall teichoic acid in *Bacillus subtilis*. Identification of genes and regulation. *J Biol Chem.*, 270(26):15598–606, Jun 1995.
- M. Perego, C. F. Higgins, S. R. Pearce, M. P. Gallagher, and J. A. Hoch. The oligopeptide transport system of *Bacillus subtilis* plays a role in the initiation of sporulation. *Mol Microbiol*, 5(1):173–185, Jan 1991.
- B. Persson and P. Argos. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J Mol Biol*, 237(2):182–192, Mar 1994.
- F. Piazza, P. Tortosa, and D. Dubnau. Mutational analysis and membrane topology of ComP, a quorum-sensing histidine kinase of *Bacillus subtilis* controlling competence development. *J Bacteriol*, 181(15):4540–4548, Aug 1999.
- E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and

- R. Lopez. InterProScan: protein domains identifier. *Nucleic Acids Res*, 33(Web Server issue):W116–W120, Jul 2005.
- J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schächter, Y. Chemama, A. Labigne, and P. Legrain. The protein-protein interaction map of helicobacter pylori. *Nature*, 409(6817):211–215, Jan 2001.
- D. A. Rasko, J. Ravel, O. A. Økstad, E. Helgason, R. Z. Cer, L. Jiang, K. A. Shores, D. E. Fouts, N. J. Tourasse, S. V. Angiuoli, J. Kolonay, W. C. Nelson, A. Kolstø, C. M. Fraser, and T. D. Read. The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. *Nucleic Acids Res*, 32(3):977–988, 2004.
- D. A. Rasko, M. J. Rosovitz, O. A. Økstad, D. E. Fouts, L. Jiang, R. Z. Cer, A. Kolstø, S. R. Gill, and J. Ravel. Complete sequence analysis of novel plasmids from emetic and periodontal *Bacillus cereus* isolates reveals a common evolutionary history among the *B. cereus*-group plasmids, including *Bacillus anthracis* pXO1. *J Bacteriol*, 189(1):52–64, Jan 2007.
- T. D. Read, S. N. Peterson, N. Tourasse, L. W. Baillie, I. T. Paulsen, K. E. Nelson, H. Tettelin, D. E. Fouts, J. A. Eisen, S. R. Gill, E. K. Holtzapple, O. A. Økstad, E. Helgason, J. Rillstone, M. Wu, J. F. Kolonay, M. J. Beanan, R. J. Dodson, L. M. Brinkac, M. Gwinn, R. T. DeBoy, R. Madpu, S. C. Daugherty, A. S. Durkin, D. H. Haft, W. C. Nelson, J. D. Peterson, M. Pop, H. M. Khouri, D. Radune, J. L. Benton, Y. Mahamoud, L. Jiang, I. R. Hance, J. F. Weidman, K. J. Berry, R. D. Plaut, A. M. Wolf, K. L. Watkins, W. C. Nierman, A. Hazen, R. Cline, C. Redmond, J. E. Thwaite, O. White, S. L. Salzberg, B. Thomason, A. M. Friedlander, T. M. Koehler, P. C. Hanna, A. Kolstø, and C. M. Fraser. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature*, 423(6935):81–86, May 2003.
- J. Reizer, S. Bachem, A. Reizer, M. Arnaud, M. H. Saier, and J. Stülke. Novel phos-

- phototransferase system genes revealed by genome analysis - the complete complement of PTS proteins encoded within the genome of *Bacillus subtilis*. *Microbiology*, 145 (Pt 12):3419–3429, Dec 1999.
- M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052, Dec 2001.
- M. W. Rey, P. Ramaiya, B. A. Nelson, S. D. Brody-Karpin, E. J. Zaretsky, M. Tang, A. L. de Leon, H. Xiang, V. Gusti, I. G. Clausen, P. B. Olsen, M. D. Rasmussen, J. T. Andersen, P. L. Jørgensen, T. S. Larsen, A. Sorokin, A. Bolotin, A. Lapidus, N. Gilleron, S. D. Ehrlich, and R. M. Berka. Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species. *Genome Biol*, 5(10):R77, 2004.
- M. Rezwan, T. Grau, A. Tschumi, and P. Sander. Lipoprotein synthesis in mycobacteria. *Microbiology*, 153(Pt 3):652–658, Mar 2007.
- D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23(8):951–959, Aug 2005.
- M. I. Roncero. Genes controlling xylan utilization by *Bacillus subtilis*. *J Bacteriol*, 156(1):257–263, Oct 1983.
- G. A. Rufo, B. J. Sullivan, A. Sloma, and J. Pero. Isolation and characterization of a novel extracellular metalloprotease from *Bacillus subtilis*. *J Bacteriol*, 172(2):1019–1023, Feb 1990.
- A. G. Rust, E. Mongin, and E. Birney. Genome annotation techniques: new approaches and challenges. *Drug Discov Today*, 7(11 Suppl):S70–S76, Jun 2002.
- S. Bornemann and group. Microbes in Norwich: *Bacillus subtilis*. <http://www.micron.ac.uk/organisms/bsu.html>. Accessed 12 Apr 2007.

- I. Sá-Nogueira, T. V. Nogueira, S. Soares, and H. de Lencastre. The *Bacillus subtilis* L-arabinose (ara) operon: nucleotide sequence, genetic organization and expression. *Microbiology*, 143 (Pt 3):957–969, Mar 1997.
- M. Saarela, G. Mogensen, R. Fondn, J. Mättö, and T. Mattila-Sandholm. Probiotic bacteria: safety, functional and technological properties. *J Biotechnol*, 84(3):197–215, Dec 2000.
- M. Sára and U. B. Sleytr. S-layer proteins. *J Bacteriol*, 182(4):859–868, Feb 2000.
- W. Schumann, G. Homuth, and A. Mogk. The GroE chaperonin machine is the major modulator of the CIRCE heat shock regulon of *Bacillus subtilis*. *J Biosci*, 23:415–422, 1998.
- M. Senger, P. Rice, and T. Oinn. Soaplab - a unified sesame door to analysis tools. In S. J. Cox, editor, *Proceedings UK e-Science, All Hands Meeting 2003*, pages 509–513, Sept 2003.
- P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.
- R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*, 102(6):1974–1979, Feb 2005.
- M. J. J. B. Sibbald, A. K. Ziebandt, S. Engelmann, M. Hecker, A. de Jong, H. J. M. Harmsen, G. C. Raangs, I. Stokroos, J. P. Arends, J. Y. F. Dubois, and J. M. van Dijk. Mapping the pathways to Staphylococcal pathogenesis by comparative secretomics. *Microbiol Mol Biol Rev*, 70(3):755–788, Sep 2006.
- A. Sloma, A. Ally, D. Ally, and J. Pero. Gene encoding a minor extracellular protease in *Bacillus subtilis*. *J Bacteriol*, 170(12):5557–5563, Dec 1988.

- A. Sloma, G. A. Rufo, C. F. Rudolph, B. J. Sullivan, K. A. Theriault, and J. Pero. Bacillopeptidase F of *Bacillus subtilis*: purification of the protein and cloning of the gene. *J Bacteriol*, 172(3):1470–1477, Mar 1990.
- Soaplab. URL <http://www.ebi.ac.uk/Tools/webservices/soaplab/overview>. Accessed 2 May 2007.
- A. L. Sonenshein, J. A. Hoch, and R. Losick, editors. *Bacillus Subtilis and Its Closest Relatives: From Genes to Cells*. ASM Press, 2002.
- M. Soriano, P. Diaz, and F. I. J. Pastor. Pectate lyase c from *bacillus subtilis*: a novel endo-cleaving enzyme with activity on highly methylated pectin. *Microbiology*, 152 (Pt 3):617–625, Mar 2006.
- D. Srivastava and Y. Velegrakis. Intensional associations between data and metadata. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 401–412, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-686-8.
- B. J. Stapley and G. Benoit. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pac Symp Biocomput*, pages 529–540, 2000.
- C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–D539, Jan 2006.
- R. Stevens, C. Goble, P. Baker, and A. Brass. A classification of tasks in bioinformatics. *Bioinformatics*, 17(2):180–188, Feb 2001.
- R. Stevens, R. McEntire, C. Goble, M. Greenwood, J. Zhao, A. Wipat, and P. Li. myGrid and the drug discovery process. *Drug Discovery Today: BIOSILICO*, 2(4): 140–148, Jul 2004a.

- R. Stevens, J. Zhao, and C. Goble. Using provenance to manage knowledge of in silico experiments. *Brief Bioinform*, 8(3):183–194, May 2007.
- R. D. Stevens, A. J. Robinson, and C. A. Goble. myGrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19 Suppl 1:i302–i304, 2003.
- R. D. Stevens, H. J. Tipney, C. J. Wroe, T. M. Oinn, M. Senger, P. W. Lord, C. A. Goble, A. Brass, and M. Tassabehji. Exploring Williams-Beuren syndrome using myGrid. *Bioinformatics*, 20 Suppl 1:I303–I310, Aug 2004b.
- C. E. V. Storm and E. L. L. Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99, Jan 2002.
- G. Subramanian, R. Mural, S. L. Hoffman, J. C. Venter, and S. Broder. Microbial disease in humans: a genomic perspective. *Mol Diagn*, 6(4):243–252, Dec 2001.
- J. W. Suh, S. A. Boylan, S. M. Thomas, K. M. Dolan, D. B. Oliver, and C. W. Price. Isolation of a secY homologue from *Bacillus subtilis*: evidence for a common protein export pathway in eubacteria. *Mol Microbiol*, 4(2):305–314, Feb 1990.
- S. Sun, Y. Zhao, Y. Jiao, Y. Yin, L. Cai, Y. Zhang, H. Lu, R. Chen, and D. Bu. Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm. *FEBS Lett*, 580(7):1891–1896, Mar 2006.
- Y. Sun, A. Wipat, M. Pocock, P. A. Lee, P. Watson, K. Flanagan, and J. T. Worthington. A Grid-based system for microbial genome comparison and analysis. In *5th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2005)*, Cardiff, UK, May 2005.
- I. C. Sutcliffe and R. R. B. Russell. Lipoproteins of Gram-positive bacteria. *J Bacteriol*, 177(5):1123–1128, Mar 1995.
- K. D. Swenson. Workflow and Web service standards. *Business Process Management Journal*, 11(3):218–223, 2005.

- H. Takami, C. G. Han, Y. Takaki, and E. Ohtsubo. Identification and distribution of new insertion sequences in the genome of alkaliphilic *Bacillus halodurans* C-125. *J Bacteriol*, 183(14):4345–4356, Jul 2001.
- H. Takami, K. Nakasone, Y. Takaki, G. Maeno, R. Sasaki, N. Masui, F. Fuji, C. Hirama, Y. Nakamura, N. Ogasawara, S. Kuhara, and K. Horikoshi. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res*, 28(21):4317–4331, Nov 2000.
- R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.
- Taverna Project Website. URL <http://taverna.sourceforge.net/index.php>. Accessed 1 May 2007.
- J. Taylor. National e-Science Centre. URL <http://www.nesc.ac.uk/nesc/define.htmls>. Accessed 10 Jun 2007.
- H. Tettelin, V. Massignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*, 102(39):13950–13955, Sep 2005.

- H. Tjalsma, H. Antelmann, J. D. H. Jongbloed, P. G. Braun, E. Darmon, R. Dorenbos, J. F. Dubois, H. Westers, G. Zanen, W. J. Quax, O. P. Kuipers, S. Bron, M. Hecker, and J. M. van Dijl. Proteomics of protein secretion by *Bacillus subtilis*: separating the "secrets" of the secretome. *Microbiol Mol Biol Rev*, 68(2):207–233, Jun 2004.
- H. Tjalsma, A. Bolhuis, J. D. Jongbloed, S. Bron, and J. M. van Dijl. Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. *Microbiol Mol Biol Rev*, 64(3):515–547, Sep 2000.
- R. J. Turner, Y. Lu, and R. L. Switzer. Regulation of the *Bacillus subtilis* pyrimidine biosynthetic (pyr) gene cluster by an autogenous transcriptional attenuation mechanism. *J Bacteriol*, 176(12):3708–3722, Jun 1994.
- U.S. Environmental Protection Agency. *Bacillus subtilis* final risk assessment. URL <http://www.epa.gov/oppt/biotech/pubs/fra/fra009.htm>. Accessed 29 Jan 2007.
- U.S. Environmental Protection Agency. *Bacillus licheniformis* final risk assessment. Accessed 7 Sept 07. URL <http://www.epa.gov/oppt/biotech/pubs/fra/fra005.htm>.
- S. van Dongen. MCL - a cluster algorithm for graphs. URL <http://micans.org/mcl/>. Accessed 27 Jan 2006.
- S. van Dongen. A cluster algorithm for graphs. Technical report ins-r0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, 2000a. Accessed 10 May 2007.
- S. van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, May 2000b.
- K. H. van Wely, J. Swaving, R. Freudl, and A. J. Driessen. Translocation of proteins across the cell envelope of gram-positive bacteria. *FEMS Microbiol Rev*, 25(4):437–454, Aug 2001.

- A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*, 21(6):697–700, Jun 2003.
- B. Veith, C. Herzberg, S. Steckel, J. Feesche, K. H. Maurer, P. Ehrenreich, S. Bumer, A. Henne, H. Liesegang, R. Merkl, A. Ehrenreich, and G. Gottschalk. The complete genome sequence of *Bacillus licheniformis* DSM13, an organism with great industrial potential. *J Mol Microbiol Biotechnol*, 7(4):204–211, 2004.
- J. Vlasblom, S. Wu, S. Pu, M. Superina, G. Liu, C. Orsi, and S. J. Wodak. Genepro: a cytoscape plug-in for advanced visualization and analysis of interaction networks. *Bioinformatics*, 22(17):2178–2179, Sep 2006.
- C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krger, B. Snel, and P. Bork. STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):D358–D362, Jan 2007.
- W3C. About the World Wide Web. URL <http://www.w3.org/WWW/>. Accessed 16 Jun 2007.
- W3C. Web services architecture. 2004. URL <http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/>. Accessed 16 Jun 2007.
- E. Wahlström, M. Vitikainen, V. P. Kontinen, and M. Sarvas. The extracytoplasmic folding factor prsa is required for protein secretion only in the presence of the cell wall in *bacillus subtilis*. *Microbiology*, 149(Pt 3):569–577, Mar 2003.
- D. P. Wall, H. B. Fraser, and A. E. Hirsh. Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1711, Sep 2003.
- H. Wang, J. Z. Huang, Y. Qu, and J. Xie. Web services: problems and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(3):309–320, Apr 2004.

- D. V. Ward, O. Draper, J. R. Zupan, and P. C. Zambryski. Peptide linkage mapping of the *Agrobacterium tumefaciens* vir-encoded type iv secretion system reveals protein subassemblies. *Proc Natl Acad Sci U S A*, 99(17):11493–11500, Aug 2002.
- D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998. URL <http://dx.doi.org/10.1038/30918>.
- M. S. Wilke, A. L. Lovering, and N. C. J. Strynadka. Beta-lactam antibiotic resistance: a current structural perspective. *Curr Opin Microbiol*, 8(5):525–533, Oct 2005.
- R. C. Williams. *A comparative analysis of the Bacillus subtilis and Bacillus anthracis secretomes*. PhD thesis, Cell and molecular biology, Nov 2002.
- L. Winstedt, K. Yoshida, Y. Fujita, and C. von Wachenfeldt. Cytochrome bd biosynthesis in *Bacillus subtilis*: characterization of the cydABCD operon. *J Bacteriol*, 180(24):6571–6580, Dec 1998.
- A. Wipat and C. R. Harwood. The *Bacillus subtilis* genome sequence: the molecular blueprint of a soil bacterium. *FEMS Microbiology Ecology*, 28(1):1–9, 1999.
- K. Woodson and K. M. Devine. Analysis of a ribose transport operon from *Bacillus subtilis*. *Microbiology*, 140 (Pt 8):1829–1838, Aug 1994.
- L. Wu and N. E. Welker. Cloning and characterization of a glutamine transport operon of *Bacillus stearothermophilus* NUB36: effect of temperature on regulation of transcription. *J Bacteriol*, 173(15):4877–4888, Aug 1991.
- X. Wu, L. Zhu, J. Guo, D. Zhang, and K. Lin. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res*, 34(7):2137–2150, 2006.
- X. C. Wu, S. Nathoo, A. S. Pang, T. Carne, and S. L. Wong. Cloning, genetic organization, and characterization of a structural gene encoding bacillopeptidase F from *Bacillus subtilis*. *J Biol Chem*, 265(12):6845–6850, Apr 1990.

- K. Xia, D. Dong, and J. J. Han. IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, 7:508, 2006.
- D. Xu and J. Côté. Phylogenetic relationships between *Bacillus* species and related genera inferred from comparison of 3' end 16S rDNA and 5' end 16S-23S ITS nucleotide sequences. *Int J Syst Evol Microbiol*, 53(Pt 3):695–704, May 2003.
- K. Xu and M. A. Strauch. Identification, sequence, and expression of the gene encoding gamma-glutamyltranspeptidase in *Bacillus subtilis*. *J Bacteriol*, 178(14): 4319–4322, Jul 1996.
- H. Yu, X. Zhu, D. Greenbaum, J. Karro, and M. Gerstein. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res*, 32(1):328–337, 2004.
- J. Zhao, R. Stevens, C. Wroe, M. Greenwood, and C. Goble. The origin and history of in silico experiments. In *UK e-Science All Hands Meeting*, 2004.
- C. M. Zmasek and S. R. Eddy. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14, May 2002.
- M. zur Muehlen, J. V. Nickerson, and K. D. Swenson. Developing Web services choreography standards: the case of REST vs. SOAP. *Decis. Support Syst.*, 40(1): 9–29, 2005.

Appendix A

Putative *B. subtilis* secreted proteins

Table A.1: Putative SpI proteins predicted by the BaSPP classification workflow, compared to Tjalsma et al. [2000]. Red boxes indicate proteins predicted to be SpI by BaSPP but not predicted to be secreted by Tjalsma et al. [2000]. Green boxes indicate proteins predicted to be SpI by BaSPP, but predicted to be SpII by Tjalsma et al. [2000].

Gene	Signal Peptide	SPase I
abnA/BSU28810	MKKKKTWKRFLHFSSAALAAGLIFTSAAPA EAA	
albB/BSU37380	MSPAQRRIILLYLSFIFVIGAVVYFVKS DYL	
amyE/BSU03040	MFAKRFKTSLLPLFAGFLLLFHLVLAGPAAASA ETA	
aprE/BSU10300	MRSKKLWISLLFALTIFTMAFSNMSVQA AGK	
aspB/BSU22370	MKLAKRVSALTPSTTLAITAKA KEL	
bdbA/BSU21460	MKKWTVLFLVLIAAAISIFYYY STG	
bdbD/BSU33480	MKKKQOSSAKFAVILTVVVVVLLAAIIVINNKTQOG NDA	
bglC/BSU18130	MKRSISIFITCLLITLLTMGGMIASPAASA AGT	
bglS/BSU39070	MPYLKRVLLLLVTGLFMSLFAVTATASA QTG	
bofC/BSU27750	MKRFTAYLLLGLCSAAVFLIGAPSRALG AEV	
bpr/BSU15300	MRKKTKNRLISSVLSTVVISSLLFPGAAGA SSK	
bsn/BSU32540	MTKKAWFLPLVCVLLISGWLAPAASASA QTT	
capA/BSU35880	MKKELSFHEKLLKLTQKQKKKTNKHVFLAIPVFLMF AFMW AGKA ETP	
cccA/BSU25190	MKWNPLIPFLLI AVLGLTFFLSYKG LDD	
comEA/BSU25590	MNWLNQHKKAII LAASAAVFTAIMIFLATGKNKEPVKQ AVP	
comGC/BSU24710	MNEKGFTLVEMLVLFISILLITIPNVTK HNQ	
comGD/BSU24700	MNIKLNEEKGFILLESLLVLSLASILLVAVFTTLPAYDNTAVR QAA	
comGG/BSU24670	MYRTRGFTYPAVL FVSALVLLIVNFVAA QYI	
cotC/BSU17700	MKNRLFILICFCVICLFLSFGQPFFPSMILTVA AKS	
cotN/BSU24620	MGMKKKLSLGVASAAALGLALVGGGTWA AFN	
csu/BSU26890	MKISMQKADFWKKAISLLVFTMFFILMMSETVFA AGL	
cwlD/BSU01530	MRKKLKWLSFLLGFIILLFLFKYQESN NDS	

cwlO/BSU34800	MRKSLITLGLASVIGTSSFLIPFTSKTASA ETL
dacB/BSU23190	MRIFKKAVFVIMISFLIATVNVNTAHA AID
dacF/BSU23480	MKRLLSTLLIGIMLLTFAPSAFA KQD
divIB/BSU15240	MNPGQDREKTVNEERIPKIKEQRKQKANRRISFIMLFFIMVLI VYLQTPISKV STI
divIC/BSU00620	MNFSRERTITEIQNDYKEQVERQNQLKKRRRKGLYRRLTVFG ALVFLTAIVLASSVWS QTS
dltD/BSU38530	MKKRFFGPILAFILFAGALA IPS
epi/BSU38400	MKNMSCKLVS SVTLFFSLTIGPLAHA QNS
ezaA/BSU29610	MEFVIGLLIVLLALFAAGYFFRKKTYA EID
fabF/BSU11340	MTKKRVVVTGLGALSPLGNDVDTSWNNAINGVSGIGPITRYDA EEY
filL/BSU16300	MKKKLMILLIILIVIGALGA AAY
ftsL/BSU15150	MSNLAYQPEKQQRHAISPEKKVTVKKRASITLGEKVLLVFAA AVLSVSLIVSKAYA AYQ
ggf/BSU18410	MKRTWNVCLTALLSVLLVAGSVPEHA EAK
glpQ/BSU02130	MRKNRILALFVLSLGLSFMVTPVSA ASK
gtaB/BSU35670	MKKVRKAIIIPAAGLGTFLPATKA MPK
ldh/BSU03050	MMNKHVNKVALIGAGFVGSSYAFALINOG ITD
licB/BSU38590	MNILLVCAAGMSTSLVSKMEKS AQE
lipA/BSU02700	MKFVKRRRIALVTILMLSVTSLFALQPSAKA AEH
lipB/BSU08350	MKKVLMAFIICLSLILSVLAAPPSSGAKA ESV
lytB/BSU35630	MKSCCKQLVCSLAAILLIPSVSEA ADS
lytC/BSU35620	MRSYIKVLTMCFLGLILFVPTALA DNS
lytD/BSU35780	MKKRLIAPMLLSAASLAFFAMSGSAQA AAY
lytE/BSU09420	MKKQIITATTAVVLGALFAHTSIR ELK
lytF/BSU09370	MKKKLAAGLTASAIVGTTLVTPAEA ATI
lytG/BSU31120	MARKKLKKRKLILISLFLVSIPLALFVLA TTL
lytH/BSU32340	MITDIFKPGCRKLCVFNMKGDYFVKVLLSALLLLFAFEPSSAG KKL
lytR/BSU35650	MRNERRKKKKTLLLTILTIIGLLVLGTGGYA YYL
mdh/BSU29120	MGNTRKKVSVIGAGFTGATTAFLLAQKELA DVV
mpr/BSU02240	MKLVPFRKQWFAYLTVLCLALAAAVSFGVPAKA AEN
mreC/BSU28020	MPNKRLMLLLLCIILVAMIGFSLKGGRR NTT
mreE/BSU31640	MASQILLNVFLAFGWMFLSNSPSAAG FIT
mtbP/BSU20250	MSKLRVMSLFSGIGAFEA ALR
nadB/BSU27870	MSKKTIAVIGSGAAALSLAAA FPP
nprB/BSU11100	MRNLTSTLLLAGLCTAAQMVFVTHASA EES
nprE/BSU14700	MGLGKKLSVAVAASFMSLSISLPGVQA AEG
ntdC/BSU10530	MKKIGIIGAGGIARAHA TAL
nucA/BSU03430	MTTDIIKTILLVTIILAAAAGVLIKGDFFES ADQ
nucB/BSU25750	MKKWMAGLFLAAAVLLCLMVPOQIOG ASS
orfRM1/BSU19590	MKRQLKLFVILITAVVASALTFLITGNSSILGQKSASTG DSK
pbp/BSU18350	MKKSIKLYVAVLLLFVVASVPYMHQAALA AEK
pbpB/BSU15160	MIQMPKKNKFMNRGAAILSICFALFFVILGRMAYITQITGA NGE
pbpD/BSU31490	MTMLRKIIIGWILLCCIPLFAFTVIASG KEV
pbpF/BSU10110	MFKIKKKKLFIPILVLTAFLALIGYISIFL GHY
pbpX/BSU16950	MTSPTRRTAKRRRRKLNKRGLLFLGLAVMVCITWNLHR NSEE NEP
pdaA/BSU07980	MKWMCSICCAAVLLAGGAQA EAV
pdaB/BSU01570	MNHFYVWHIKRVKQIILIAFAAAASFFYIQRVPLPVFS TDT
pel/BSU07560	MKKVMLATALFLGLTPAGANA ADL
pelB/BSU18650	MKRLCLWFTVFSFLVLLPGKALG AVD
penP/BSU18800	MKLKTKASIKFGICVGLLCLITGFTPFNFSTHAEA KSI

pgdS/BSU35860	MNITLANWKKFLLVAVIICFLVPIMTKAEIAEA DTS
phoA/BSU09410	MKKMSLFQNMKSKLLPIAAVSVLTAGIFAGAELQOTEKA SAK
phoB/BSU05740	MKKFPKLLPIAVLSSIAFSSSLASGSVPEASA QEK
phrA/BSU12440	MKSKWMSGLLLVAVGFSFTQVMVHA GET
phrC/BSU03780	MKLKSKLFVICLAAAIFTAAGVSANA EAL
phrE/BSU25840	MKSKLFISLSAVLIGLAFFGSMYNGEMK EAS
phrF/BSU37470	MKLKSKLLLSCLALSTVFVATTIANA PTH
phrG/BSU40310	MKRFLIGAGVAAVILSGWFIA DHQ
phrI/BSU05020	MKISRILLAAILSSVFSITYLOS DHN
phrK/BSU18920	MKKLVLCVSLAVILSGVALTQSTSPSNIOV AER
phy/BSU19800	MKVPKTMILLSTAAGLLSLTATSVSA HYV
sacB/BSU34450	MNIKKFAKQATVLTFTTALLAGGATQAF KET
sacC/BSU27030	MKKRLIQVMIMFTLLLTMAFSADA ADS
sdhA/BSU28440	MSQSSIVVGGGLAGLMATIKAE AES
sipU/BSU04010	MNAKTITLKKRKKIKTIVVLSIIMIAALIFTIRLVFYKPFLEIGSS MA PTL
sipV/BSU10490	MKKRFWFLAGVVSIVLAJOV KNA
sleB/BSU22930	MKSKGSIMACILFSFTITFTETISA FSN
spoIID/BSU36750	MKQFAITLSVLCALILLVPTLLVIPFQHNKEAG ASV
spoIIAB/BSU24420	MLKLLGAVFIVVATTWTGFEMA KTY
spoIIAH/BSU24360	MLKKQTVWLLTMLSLVVVLSVYYTMSPEKNAVQMQSEKS ASD
spoIIP/BSU25530	MRNKRNRNQIVVAVNGGKAVKAIFLFTVSLIVFVLSGVLTSLR PEL
spoIIR/BSU36970	MKKTVICTYIFLLLSGALVGLAKEETA QKS
spoVD/BSU15170	MRVSNVTVRKRLFLVLLFGVTVFLIIDTRLGYVQFVMGEKLTSL AKDS
sppA/BSU29530	MNAKRWIALVIALGIFGVSIIVSISMSFFESYKG AQT
stoA/BSU13840	MLTKRLLITYIMLLGLIAWFGAAQA EEK
tatAc/BSU17710	MELSF TKILVILFVGLVFGPDKLPALGRAAG KAL
tatAd/BSU02630	MFSNIGPGLILIFVIAIIFGPKLPEIGRA AKR
tatAy/BSU05980	MPIGPGSLAVIAIIVALIIFGPKKLPGLGA AGD
vpr/BSU38090	MKKGIIRFLLVSFVLFALSTGITGVQA APA
wapA/BSU39230	MKKRKRNRNFKRFIAAFLVLAALMISLVPADVLA KST
wprA/BSU10770	MKRRKFSSVVAAVLIFALIFSLFSPGKAAA AGA
xpaC/BSU00250	MQRFFHFLVWSLTSSATFVFIGILSFFGLNQSIFLSIVYGLASGA AVY
xynA/BSU18840	MFKFKKNFLVGLSAAALMSISLFSATASA AST
yacD/BSU00720	MKSRTIWTIILGALLVCCIAVAYTLTKSQAGA SSS
ybbC/BSU01650	MRKTIFAFLTGLMMFGTITAASA SPD
ybbE/BSU01670	MKTKTILFIFSAITLSIFAPNETEA QTA
ybbR/BSU01760	MDKFLNRRWAVKIIALLFALLLYVAVNSNO APT
ybcS/BSU01930	MNSLSLVFWSILAVVGLLLFIKFKPPTIASLLSKDEA KEI
ybdG/BSU01990	MKTLWKVLKIVFVSLAALVLLVSYSY FIY
ybdN/BSU02040	MVKKWLIQFAVMSVLSFTYSASAVGYTA ITG
ybfO/BSU02310	MKRMIVRMTLPLLIVCLAFSSFSASARA ASE
ybxI/BSU02090	MKKWTVVVLVLSIAGIGGFSVHA ASS
ybyB/BSU02110	MKQKLLLSGLAVSTVGITSYLLKDPNSR QKA
ybcB/BSU02460	MSRIRKAPAGILGFPVAPFNTQGTLEE EAL
yckD/BSU03400	MKRITINITMFIAAAVISLTGTAEAE AEK
yelB/BSU03630	MKAEFKRKGGGKVKLVVGMGTATGATGVRLLQWLKA AGV
ydbE/BSU04440	MKSLLAALMIAGIATATLFIG FHD
ydcC/BSU04630	MRLLYASVLPKRVPSVPLKSNKRFLPFTSIEKKGLKKVRKSF VLLLTGLLAVLILSACG QKT
yddH/BSU04970	MISKKVVLPLVFSAPFIFVLCIVVVMITSR ENQ
yddl/BSU04980	MKTHIAWASACLLVMLTGFTTIGQ QTY

yddT/BSU05100 MRKKRVITCVMAASLTGSLLPAGYASA KED
 ydhH/BSU05760 MSSHYKYPLIFTAFLIAFCLIFFS YHL
 ydhM/BSU05810 MKKILLACSSGMSTSLVTKMKEYA QSI
 ydhT/BSU05880 MFKKHTISLLIIFLLASAVLAKPIEA HTV
 ydiK/BSU06000 MRNPVVWGMIFYAVGCIITYLAASSPGSMWSFYSLLMVFA
 AYN
 ydjM/BSU06250 MLKKVILAAFILVGSTLGAFSFSDDASA KHV
 ydjN/BSU06260 MKKRILLAVTIAAAAAAGVAFY VAK
 yerD/BSU06590 METIIIALIAFIIGIIAIPVLF A WTY
 yesW/BSU07050 MRRSCLMIRRRKRMTAVTLLVLLVMGTSVCPVKA EGA
 yetM/BSU07230 MKHMLIAGGGIGGLSAAISLRKAGFS VTL
 yfhK/BSU08570 MKKKQVMLALTAAAGLGLTALHSAPAAKA APL
 yfkD/BSU07930 MMKKLFHSTLIVLLFFSFFGVQPIHA KKQ
 yfkI/BSU07890 MNNERLMLKGIFLGAAAGAALLLHKPTRQ ACG
 ygaI/BSU08760 MSFFKKLAASAGIGA AKV
 yhaH/BSU10000 MASGRSLLTGLFVGIGGAAVLLTAPSSG KQL
 yhaK/BSU09960 MRTWKRIPTTMLSLSVSPFLITPVLEVA ALA
 yhaL/BSU09940 MLFFPWVWVYLCIVGIISAYKLVAAA KEE
 yhcC/BSU09030 MAIIIAIAAVTVIAALITENVRNA SPG
 yhcM/BSU09140 MLFNQRRGISPAALIGSTMLITALSPQIROR ISG
 yhcO/BSU09160 MLAVSSAIVSSAMYLSFPGQASG ITK
 yhdC/BSU09360 MKSLPYTIALLCGLIIVSMAAKGHS TDT
 yheN/BSU09660 MRQTRGEASPSAVSLAFKFASLAVLCVLLLLMVLGYNSSTK
 A KEV
 yhmM/BSU10280 MKKIVAAIVVIGLVFIAFFYLYSRSGDYQ SVD
 yhjA/BSU10440 MKKAAAVLLSLGLVFGFSYGAGHVAEA KTK
 yhjC/BSU10460 MKLIHVLAALPFIGILLGIPFANK VTP
 yitP/BSU11070 MQKSISFFVIFSILWGSFLFSIIGSLGTTPIPLTK DSK
 yitY/BSU11170 MKKKLLAFALCTGAYAAALFAYSVNS EQK
 yjcM/BSU11910 MKKELLASLVCLSLSPVSTNEVFA ATT
 yjcN/BSU11920 MKKKTKIILSLAALIVILVLPVLSPVVET ASS
 yjdB/BSU11990 MNFKKTIVVSALSISALASVSGVASA HEI
 yjeA/BSU12100 MLAKRIKWVHVLIIVVCVVGIGFFHNHSLK KEI
 yjfa/BSU12110 MKRLFMKASLVLFVAVVFAVKGAPAKA ETH
 yjiA/BSU12200 MAAQTDYKKQVVGILLSLAFVLVVFVSFSEK HEK
 yknX/BSU14350 MKKVWIGIGIAVIVALFVGINIRSAAPTSGS AGK
 ykoJ/BSU13280 MLKKKWMVGLLAGCLAAGGFSYNFAA TEN
 ykoQ/BSU13370 MFTILLSLAILVFPFLYKANR NTK
 ykpC/BSU14460 MLRDLGRRVAIAAILSGIILGGMSSISLANMPHSPAGG TVK
 ykuA/BSU13980 MNLFFLAVFVLTALIFKLGVVQIVEG EQH
 ykuE/BSU14050 MKKMSRRQFLKGMFGALAAGALTAGGGYGYA RYL
 ykuO/BSU14160 MFHKGATAVTASAFSGYFVAVQR EGI
 ykvT/BSU13820 MITKFTALAVFLLCFMPAAKIEHT QAS
 ykwD/BSU13970 MKKAFILSAAAAGVGLFTFGGVQQAASA KEL
 ylaE/BSU14750 MKKTFVKKAMLTAAAMTSAALLTFGPDAASA KTP
 ylaF/BSU14760 MKKMNWLLLLFAFAAVFSIMLIGVFIA EKS
 ylbC/BSU14960 MKNILRAMVILLICGTIVLFIQYGSVPEKK SND
 ylbL/BSU15050 MLRKKHFSWMLVILLIAVLVFIKLPYYITKPGA TEL
 yliqB/BSU15960 MKKIGLLFMLCLAALFTIGFPAQQADA AEA
 ylxF/BSU16260 MSGKKKESGKFRSVLLIILPLMFLLIAGGIVLWA AGI
 ylxW/BSU15250 MRGKSAVLLSLIMLIAGFLISFSFQMTKENN KSA
 ylxX/BSU15260 MKIKRSFISIVLMVIFGLMISVQFNSI KHP
 ylxY/BSU16700 MYKKFVPFAVFLFLFFVSFEMMENPHA LDY

ymaC/BSU17270	MRRFLNVLVLAVLFLRYVHY SLE
ymdA/BSU16960	MPIMMVLISILLGLVVGYPVRKTIAEA KIA
ymcM/BSU17690	MAKPLSKGGILVKKVLIAGAVGTAVLFGTLSSGIPGLPAADAQ VAKA ASE
ymdA/BSU17720	MRFTKVVGFLSVLGLAAVFLTAQA EKA
ymdL/BSU17820	MKPAKVSLRLHSLKHVDCNIAKRFPSITIKVLLMIFMVFI PISSIYA EDV
ymeA/BSU17860	MMSKESIIFVGLFTVILSAVILMLSYTSSG QEL
ymfF/BSU18150	MPRIKKITICVLLVCFTMLSVMLGPGATEVLA ASD
ymgB/BSU18180	MRKKVRKAVIPAAGLGTRFLPATKA QPK
ymgK/BSU18280	MKVCQKSVRFLVSLIIGTFVISVPPFMANA QSD
yoaD/BSU18560	MKNTMKRMFCMTVLVITAPYNEEGR KEL
yoaW/BSU18780	MKKMLMLAFTFLALITHVGEASA VIV
yobB/BSU18820	MKIRKILLSSALSFGMLISAVPALA AGT
yobV/BSU19100	MKLERLLAMVVLISKKQVOA AEL
yocA/BSU19130	MKKKRKGCFAAAGFMIFVFIASFLVLLFENR DLI
yocH/BSU19210	MKKTIMSFVAVAALSTAFGAHASA KEI
yoeb/BSU18380	MKKCLLFTTIALILSLSTNAEA KNT
yojL/BSU19410	MKKKTIVAGLAVSAVVGSSMAAAPAEA KTI
yola/BSU21540	MKKRITYSLALLAVVAFAFTDSSKAKA AEA
yolC/BSU21520	MKKRLIGFLVLPALIMSGITLIEA NKK
yomL/BSU21320	MRKKRVITCVMAASLTGSLLPAGYATA KED
yoni/BSU21000	MLEKMGIVVAFILSLTTLTINSLTIVE KVR
yopL/BSU20850	MKKLIMALVILGALGTSYISA DSS
yoqH/BSU20630	MKRFLVLSFLSIIVAYPIQTNA SPM
yoqM/BSU20580	MKLRKVLTVGSVLSLGLLVASPAFA TSP
yorM/BSU20330	MFKKLIDKHKKYVYHRINKMALFATIGLLGVGLYYS AKN
vosU/BSU20000	MRKILKIVSLILLILLVYSFSPNSQLFVYV QLI
voxC/BSU18510	MITVYISLAVLAVSIIFLGVTVIQN KKK
vozM/BSU18960	MKKRLIGFLVLPALIMWGITLIES NKK
ypbG/BSU22980	MKL SVKIAGVLTVAAMTAAMTAATA KGN
ypjP/BSU21840	MKLWMRKTLVVLFTIVTGLVSPPAALMA DKP
ypmB/BSU22380	MRKKALIFTVIFGIIFLAVLLYSA SIY
ypmS/BSU21730	MNKWKRLFFILLAINFILAAGFVALVLLPGEQA QVK
ypuA/BSU23370	MKKTWIGMLAAAVLLMVPKVSLA DAA
ypuD/BSU23300	MGRIKTKTILLVLLLLAGGYMYNDIEL KDV
yqfA/BSU25380	MDPSTLMILAIVAVAVLAVFFIFVPVMLWISALA AGV
yqfZ/BSU25060	MKRLTLVCSIVFILFILFYDLKIGITPIQDLPIYE ASA
yqgA/BSU25050	MKQGKFSVFLILLMLTLVVPKGAFA ASS
yqgF/BSU25000	MRRNPKKQNHKEKKKSLPIRLNILFLAAFVIFTWIIIVELGIKQI VQG DDY
yqgW/BSU24800	MLLVVIFGLVALFALWGVLSVR NKN
yqhP/BSU24500	MNHRVQPIIAVLIALGAFGLYVLVTNPGEMA KMA
yqjB/BSU23940	MRFFLCIFMMISPIWPLGENPLPG DPY
yqjU/BSU23740	MNL SRKHEVKNMNIGDVMFQLFVFIIFAAVYFA AVT
yqkD/BSU23640	MKKILLAIGALVTAVIAIGVFSHMLFIKK KTD
yqxA/BSU25520	MIAYFGKCLLLVTIMFLGVLFQMQQANHG MLS
yqxI/BSU25890	MFKKLLLATSLTSLVLPDGHAKA QEV
yqxM/BSU24640	MFRLFHNQKAKTKLVLLIFQLSVIFSLTAAICLQFSDDTSA FHD
yqzC/BSU24940	MTKRGIAQAFAGGIIATAVLAADVFLYLTDEDQA AAV
yqzD/BSU24930	MEIAIALFTVSIALIAFSYSOR DPM
yqzG/BSU24650	MMIKQCVICLSLLVFGTTAAHA EET

yraI/BSU26930	MVSESKSLTGCKKVKRTAFIRGGYKVNKLKRLSMLTVMIASV FIFSSHALA AQY
yraJ/BSU26920	MTLTKLKMLSMLTVMIASLFISSQALAVQYFTVSTSSG APV
yrrL/BSU27370	MYTNQQKKSFFNKRIILSSIVVLFLLIGGAFLYGKSLLEPV EK DSK
yrrR/BSU27310	MKISKRMKLAVIAFLIVFFLLRLA EIQ
yrrS/BSU27300	MSNNQSRyenRDkRRKANLVNIIIAIVSILIVVVAANLFNPS SKDVSK DSE
yrrJ/BSU27580	MNKKYFVLIVCIIFTSALFPTFSSVTA AOG EAV
ytcA/BSU30860	MKICVVGAGYVGLTLSAALA SIG
ytrI/BSU29240	MRVPQHYKKPGWQRFFAGMMCGAVISWFFLFYGTFOE EQV
yttA/BSU30360	MVLAFLGFLACIALGYGLYHLVRYVLKKEKRF SKRLFWPLFI GGLVLLFTGAALAEPTAAA NAE
ytxE/BSU29720	MKLRRERFERRNGSGKNSQSSSSWMVTFDTLITLILVFFILLFS MSQIDLOKFKA AVD
ytxG/BSU29780	MIIILYLSVALIAVAFLVLVYLSKTLKS LQL
ytxB/BSU29870	MKLRLHLLGAGLGICTAVYR QYV
yuaB/BSU31080	MKRKLSSLAISLSLGLLSAPTASFA AES
yuaG/BSU31010	MTMPIIMIGVVFFLLIALIAVFIKYRTA GPD
yufs/BSU31590	MKTkvVMCSGLFCSVFAGAFMLNOYDGRSGV AAC
yuiC/BSU32070	MLNMIRRLMTCLFLAFGTIFLSVSGIEA KDL
yunB/BSU32350	MPRYRGPFKRGRPLPRYVMLLSVVFFILSTTVSLWMING SIK
yusR/BSU32900	MKGMFFCARAVVPMKSKDGAIVNVGSIAGITGAGSSMPYA VS KSA
yusZ/BSU32980	MNKKIAIVTGASSGFLLAAVYKL ARS
yuzG/BSU32120	MAQNEEKTpkSQKIQRIMAMTWVVAALVIALVVGTA LN YIN
yvaY/BSU33770	MKSKLLRLIVSMVTILVFSVLGSKESSTSA KEN
yvbX/BSU34020	MKKWLIHAVSLAIAIVLFMYTKGEAKA AGM
yvcC/BSU34250	MILKRLFDLTAAIFLLCCTSVIILFTIAVRL KIG
yvgO/BSU33410	MKRIRIPMTLALGAALTIAPLSFASA EEN
yvhJ/BSU35520	MAERVVRVRKKKKSKRRKILKRIMLLFALALLVVVGLGGY KLYKTINA ADE
yvjB/BSU35240	MNQKIMAVIAAGSMLFGGAGVYAGINLLEMDKPQTAAPATA QA DSE
yvnB/BSU35040	MRKYTVIASILLSFLSVLSGGHHESKA FPV
yvpA/BSU34950	MKKIVSILFMFGLVMGFSQFQSTVFA ADK
yvpB/BSU34940	MKTLRLTCLVLMILSGVIFFGKIDA KDI
yvtP/BSU33280	MSKKKKWLIGGAICAGVLVLAGIGAGGFYFFTHMNQVAV SSE
ywaD/BSU38470	MKKLLTVMTMAVLTAGTLLPAQSVTPAAHA VQI
ywbN/BSU38260	MSDEQKKPEQIHRRDILKWGAMAGAAVAIGASGLGGLAPLVQ TAA KPS
ywcI/BSU38080	MKRLVSLRVWMVFLMNWVTPDRKTARA AVY
yweA/BSU37800	MLKRTSFVSSLFISSAVLLSILLPSGQAHQA QSA
ywgB/BSU37580	MKMKSGMEQAVSVLLLLSRLPVQASLTSEA ISQ
ywhE/BSU37510	MDAMTNKRLRLTLKTVRAFIFLGAFALAAA AVFMTVILIAKY QGAPSVQVPQSTILYA SDG
ywjE/BSU37190	MKVFTVIMIVVIFFALILLDIFMGRAGY RKK
ywmB/BSU36770	MKKKQVSHAIISVMSLFSVIAVFHTIHA SEL
ywmC/BSU36740	MKKRFSIMMTGLLGLTSPAF A EK
ywmD/BSU36730	MKKLLAAGIIGLLTVSIASPSFA AEK
ywmE/BSU36720	MKLFGMIFLIATVAFILLGVLLKLAFFVLSILTILIA AV

ywnE/BSU36590	MSISSILLSLFFILNILLAIIVIFKERRDA SAS
ywnI/BSU36550	MREEEKKTSQVKKLQQFFRKRWVFPATYLVSAAVILTAVLWY QSVSN DEV
ywoF/BSU36460	MRKWYFILLAGVLTSVILAFVYDKTKA NEE
ywqC/BSU36260	MGESTSLKEILSTLTRILLIMVTAAATAAGGLISFFALTPITYE NST
ywqO/BSU36140	MKFLLSVIAGLLILALYLFWKVQPPVWIOV ETN
ywsB/BSU35970	MNKPTKLFSTLALAAGMTAAAAGGAGTIHA QQP
ywtC/BSU35870	MKFVKAIWPFVAVAVFMFMSAFK FND
ywtF/BSU35840	MEERSQRRKKRKLKKWVKVVAGLMAFLVIAAGSVGAYA FVK
yxal/yxaK/BSU39940	MVKSFRMKALIAGA AVAAAVSAGAVSDVPAAKVLQPTAAYA AET
yxia/BSU39330	MFNRLFRVCFLAALIMAFILPNSYYA QKP
yxIT/BSU39030	MKWNNMLKAAGIAVLLFSVFAYAAPSLKAVQA KTP
yxjF/BSU38970	MRKQVALVTGAAGGIRFEIAREFA REG
yxzE/BSU38790	MTVIIIIFISIVFLSVIQPPSKN KSR
yybN/BSU40580	MNKFLKSNFRLLAAALGISLLASSNFIKA SND
yycH/BSU40390	MKRENIKTILLTVLVVISLVFTWGIWTFQPNFSEG SSS
yycP/BSU40270	MKKWMITIAMLILAGIALFVFISPLKSHK TV'S

Table A.2: Putative SpII proteins predicted by the BaSPP classification workflow, compared to Tjalsma et al. [2000]. Red boxes indicate proteins predicted to be SpII by BaSPP but not predicted to be secreted by Tjalsma et al. [2000]. Green boxes indicate proteins predicted to be SpII by BaSPP, but predicted to be SpI by Tjalsma et al. [2000].

Gene	Signal Peptide	SPase II
appA/BSU11380	MKRRKTALMMLSVLMVLAIFLSA <u>CS</u>	
araN/BSU28750	MKKMTVCFLVLMMLTLVLAG <u>CS</u>	
cccB/BSU35270	MKSKLSILMIGFALSVLAA <u>CG</u>	
dacA/BSU00100	MNIKKCKQLLSLVVLTAVT <u>CL</u>	
dppE/BSU12960	MKRGKRMKRVKKLWGMGLALGLSFALMG <u>CT</u>	
feuA/BSU01630	MKKISLTLLILLALTAAA <u>CG</u>	
fluD/BSU33320	MTHYKKGAAFFALLIAALAA <u>CG</u>	
gerAC/BSU33070	MKIRILCMFICTLLSG <u>CW</u>	
gerBC/BSU35820	MKTASKFSVMFFMLLALCG <u>CW</u>	
gerD/BSU01550	MSKAKTLLMSCFLLSYTA <u>CA</u>	
gerKC/BSU03710	MVRKCLLAVLMLLSVIVLP <u>CW</u>	
gerM/BSU28380	MLKKGPAVIGATCLTSALLSG <u>CG</u>	
gluH/BSU27440	MKKIFSLALISLFAVILLAA <u>CG</u>	
lplA/BSU07100	MKIRMRKKWMALPLAAMMAG <u>CS</u>	
lytA/BSU35640	MKKFIALLFFILLSG <u>CG</u>	
med/BSU11300	MITRLVMIFSVLLLLSG <u>CG</u>	
mntA/BSU30770	MRQGLMAAVLFATFALTG <u>CG</u>	
msmE/BSU30270	MKHTFVFLSLILLVLP <u>CS</u>	
oppA/BSU11430	MKKRWSIVTLMILFTLVLSA <u>CG</u>	
opuAC/BSU03000	MLKKIIGIGVSAMLALSIAA <u>CG</u>	
opuBC/BSU33710	MKRKYCLKMIGLALAATLTSG <u>CS</u>	
opuCC/BSU33810	MTKIKWLGAFAVFMVLLGG <u>CS</u>	
pbpC/BSU04140	MLKKCILVFLCVGLIGLIG <u>CS</u>	
prsA/BSU09950	MKKIAIAAITATSILALSA <u>CS</u>	
rbsB/BSU35960	MKKAVSVILTSLFLTA <u>CS</u>	
slp/BSU14620	MRYRAVFPMLIIVFALSG <u>CT</u>	
ssuA/BSU08840	MKKGLIVLVAVIFLLAG <u>CG</u>	
xynD/BSU18160	MRKKCSVCLWLVLLS <u>CL</u>	
ybbD/BSU01660	MRPVFPLLSAVLELS <u>CF</u>	
ybfJ/BSU02250	MYSTIFNIGQINKYSKLAIFMSILFLCG <u>CS</u>	
yccC/BSU02690	MKKQRMLVLTALLFVFTG <u>CS</u>	
yedA/BSU02780	MFQKKTYAVFLILLMMFTA <u>CS</u>	
yedD/BSU02810	MNLPAKTFVILCILFLDL <u>CF</u>	
yedH/BSU02850	MFKKWSGLFVIAACFLLYAA <u>CG</u>	
yelB/BSU03350	MKLSLFIIVLMPVILLSA <u>CS</u>	
yckB/BSU03380	MKSFMHKAVIFSFTMAFFLILAA <u>CS</u>	
yckK/BSU03610	MKKALLALFMVVSIAALAA <u>CG</u>	
yelQ/BSU03830	MKKFALLFIALVTAVVISA <u>CG</u>	
ydaJ/BSU04270	MRHVLIIVLFFLSIGLSAG <u>CA</u>	
yddJ/BSU04990	MKNLFIFLSLMMMFVLTAA <u>CG</u>	
ydeJ/BSU05220	MKKRRKICYCNTALLMILLAG <u>CT</u>	

ydhF/BSU05730 MRRILSILVFAIMLAG CS
 ydhK/BSU05790 MSAGKSYRKKMKQRRNMKISKYALGILMLSLVFVLSA CG
 yerB/BSU06570 MKKWMTVCALCFVFFLLVS CQ
 yerH/BSU06630 MKKTLALAATAAVLMLSA CS
 yfkR/BSU07780 MKKTIYKCVLPILLICLLTG CW
 yfmC/BSU07520 MRTYSNKLIAIMSVLLIACLIIVSG CS
 yhcJ/BSU09110 MKKWLICSFVLVLLVSFTA CS
 yhcN/BSU09150 MFGKKQVLASVLLIPLMTG CG
 yjgB/BSU12150 MKKTMSAITAAAAVTS CF
 yjha/BSU12180 MKKVLLLVLTIGLALS CA
 ykoI/BSU13270 MTKTIKTVSFAAAAILVVI CT
 vlaJ/BSU14800 MRILFIQLTLILSA CA
 yloI/BSU15700 MLNNRNVLLCVSGGIAVYKA CA
 yncB/BSU17620 MKKILISMAIVLSITLAA CG
 yndF/BSU17770 MKSKLKRQLPAMVIVCLLMICVTG CW
 yoaJ/BSU18630 MKKIMSAFVGMVLLTIF CF
 yoaO/BSU18680 MRKKNNIKKWLIIAGFLII CI
 yobA/BSU18810 MPKIGVSLIVLIMLIIFLAG CN
 yodJ/BSU19620 MKKSGKWFLAALSVTAIVGAG CS
 yojM/BSU19400 MHRLLLLMMLTALGVAG CG
 yokB/BSU21650 MNIRFSMLVCVSFIFFTGG CA
 yokF/BSU21610 MKKVLLGFAAFTLSLSLAA CS
 yonS/BSU21010 MKLFKKLGILLITSLILLAA CK
 yozF/BSU18710 MKRVLFSTVFTAVGFTF CQ
 yphF/BSU22810 MGLKCAIIFAAVFLSG CL
 ypmQ/BSU21750 MKVIKGLTAGLIFLFLCA CG
 ypmR/BSU21740 MKLRIFSIMASLILLTA CT
 yqeF/BSU25700 MKHFIILFLLFVITAG CE
 yqgG/BSU24990 MKKKNKLVLMMLMAAFMMIAAA CG
 yqgU/BSU24820 MLMRSVCFILLAVLFLSLA CK
 yqiH/BSU24200 MKQTVLLLTALFLSG CS
 yqil/BSU24190 MRMLWRSLALCGLALTAP CA
 yqiX/BSU23980 MKKWLLLLVAACITFALTA CG
 yrpD/BSU26820 MMKKGLLAGALTATVLFGT CA
 yrpE/BSU26830 MNIFSKRLGILTIGSLIVLAG CQ
 yscB/BSU28890 MNKLIQLALFFILMLTG CS
 ytcQ/BSU30160 MCLVLALGGVLAG CK
 ytkA/BSU30660 MKKMLVVLLFSALLNG CG
 ytlA/BSU30590 MNRWLRLGFACVGSIFLMFALAA CK
 ytmJ/BSU29380 MNKRKGLVLLSVFALLGGG CS
 ytmK/BSU29370 MKTKTAFMAILFSLITVLSA CG
 yunN/BSU31540 MSLVIAAGTILGA CG
 yurO/BSU32600 MKKMLFLIIAAVSMITAG CS
 yusA/BSU32730 MKKLFLGALLVFAGVMAA CG
 yusW/BSU32950 MHLIRAAGAVCLAVVLIAG CR
 yutC/BSU32320 MKRTAVSLCLLTGLLSG CG
 yvdG/BSU34610 MVLLKKGFALAAFLAIGLAA CS
 yvfK/BSU34160 MKMAKKCSVFMLCAAVSLSLAA CG
 yvfo/BSU34120 MKSKVKMFFAAAVVWSA CS
 yvgL/BSU33380 MFKKYSIFIAALTAFLVAG CS
 yvtA/BSU34850 MKKIIFICFSLILALTGG CS
 yvtC/BSU33180 MKKRAGIWAALLAAVMLAG CG
 ywbM/BSU38270 MNFTKIAVSAGCILALCAG CG
 yxeA/BSU39620 MKKAMAILAVLAAAIVI CG
 yxeB/BSU39610 MKKNILLVGMVLLLMFVSA CS
 yxeF/BSU39570 MVPLRNKYGILFLIAVCMVSG CQ
 yxeM/BSU39500 MKMKKWTVLVVAALLAVLSA CG

ysiM/BSU39120	MKKWMAAVFVMMML ML CF
ysiP/BSU39090	MRRIGLCISLLVTVLVMSA CE
yskH/BSU38800	MKRLFLSIFLLGSCLALAA CA
ysbP/BSU40560	MKIILTVLAGVGLLSAGG CG
yscO/BSU40280	MKLKKRVSMFLVALTM CG
yscS/BSU40240	MRFRWVWLFVIMLLAE CQ

Gene	Comment
<i>citH</i>	Not in proteome analysed by BaSPP.
<i>fliZ</i>	Transmembrane domain detected.
<i>mdr</i>	Transmembrane domain detected.
<i>motB</i>	No SpI/SpII detected by BaSPP.
<i>rmpG</i>	Not in proteome analysed by BaSPP.
<i>tyrA</i>	No SpI/SpII detected by BaSPP.
<i>ydbK</i>	Transmembrane domain detected.
<i>yfiS</i>	Not in proteome analysed by BaSPP.
<i>ykuV</i>	Not in proteome analysed by BaSPP.
<i>ynzA</i>	Not in proteome analysed by BaSPP.
<i>yodV</i>	No SpI/SpII detected by BaSPP.
<i>yolI</i>	No SpI/SpII detected by BaSPP.
<i>ypcP</i>	No SpI/SpII detected by BaSPP.
<i>yunA</i>	Not in proteome analysed by BaSPP.
<i>yurI</i>	Not in proteome analysed by BaSPP.
<i>yvcE</i>	Not in proteome analysed by BaSPP.
<i>yveB</i>	No SpI/SpII detected by BaSPP.
<i>yvgV</i>	Not in proteome analysed by BaSPP.
<i>ywdK</i>	Transmembrane domain detected.
<i>ywfM</i>	Transmembrane domain detected.
<i>ywtD</i>	Not in proteome analysed by BaSPP.
<i>yyaB</i>	Transmembrane domain detected.

Table A.3: SpI proteins identified by Tjalsma et al. [2000] but not BaSPP.

Gene	Comment
<i>ctaC</i>	Transmembrane domain detected.
<i>qorA</i>	Transmembrane domain detected.
<i>spoIIIJ</i>	Not in proteome analysed by BaSPP.
<i>spoIVB</i>	No SpI/SpII detected by BaSPP.
<i>ygbA</i>	Not in proteome analysed by BaSPP.
<i>yhaR</i>	No SpI/SpII detected by BaSPP.
<i>yhfQ</i>	No SpI/SpII detected by BaSPP.
<i>ykuH</i>	Transmembrane domain detected.
<i>ymzC</i>	No SpI/SpII detected by BaSPP.
<i>yqjG</i>	Not in proteome analysed by BaSPP.
<i>ytgA</i>	Not in proteome analysed by BaSPP.
<i>ytrF</i>	Transmembrane domain detected.
<i>yvcA</i>	Not in proteome analysed by BaSPP.
<i>ywnJ</i>	Transmembrane domain detected.
<i>yybM</i>	No SpI/SpII detected by BaSPP.

Table A.4: SpII proteins identified by Tjalsma et al. [2000] but not BaSPP.

Appendix B

SubtiList classification codes

1. Cell envelope and cellular processes

- (a) Cell wall
- (b) Transport/binding proteins and lipoproteins
- (c) Sensors (signal transduction)
- (d) Membrane bioenergetics (electron transport chain and ATP synthase)
- (e) Mobility and chemotaxis
- (f) Protein secretion
- (g) Cell division
- (h) Sporulation
- (i) Germination
- (j) Transformation/competence

2. Intermediary metabolism

- (a) Metabolism of carbohydrates and related molecules
 - i. Specific pathways
 - ii. Main glycolytic pathways
 - iii. TCA cycle
- (b) Metabolism of amino acids and related molecules
- (c) Metabolism of nucleotides and nucleic acids

- (d) Metabolism of lipids
- (e) Metabolism of coenzymes and prosthetic groups
- (f) Metabolism of phosphate
- (g) Metabolism of sulfur

3. Information pathways

- (a) DNA replication
- (b) DNA restriction/modification and repair
- (c) DNA recombination
- (d) DNA packaging and segregation
- (e) RNA synthesis
 - i. Initiation
 - ii. Regulation
 - iii. Elongation
 - iv. Termination
- (f) RNA modification
- (g) Protein synthesis
 - i. Ribosomal proteins
 - ii. Aminoacyl-tRNA synthetases
 - iii. Initiation
 - iv. Elongation
 - v. Termination
- (h) Protein modification
- (i) Protein folding

4. Other functions

- (a) Adaptation to atypical conditions
- (b) Detoxification

- (c) Antibiotic production
 - (d) Phage-related functions
 - (e) Transposon and IS
 - (f) Miscellaneous
5. Similar to unknown proteins
- (a) From *B. subtilis*
 - (b) From other organisms
6. No similarity

Appendix C

SubtilNet GO level threshold deduction based on *B. subtilis*

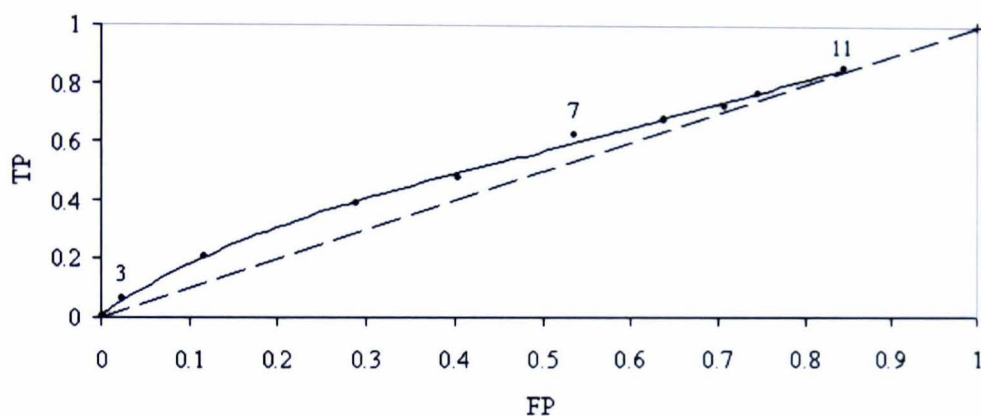


Figure C.1: Predictive capability of GO process annotations at different depth thresholds based on KEGG pathway. Points from left to right represent a threshold of depth 3 (i.e. terms at level 3 in the GO DAG and below, where root is level 1) to a maximum threshold of depth 11.

Appendix D

SubtilNet log likelihood calculations for *B. subtilis*

	Before	After	
		I = 2	I = 3
LLS	1.829	1.7700	1.8011

Table D.1: Log likelihood scores for unweighted BLASTp pairs before and after clustering with MCL, using different inflation (I) values.

	DBTBS operon	COGs	BLASTp	PREDICTOME phylogenetic profiling
LLS	3.7839	0.7568	1.8291	3.0757

Table D.2: Log likelihood scores for unweighted datasets.

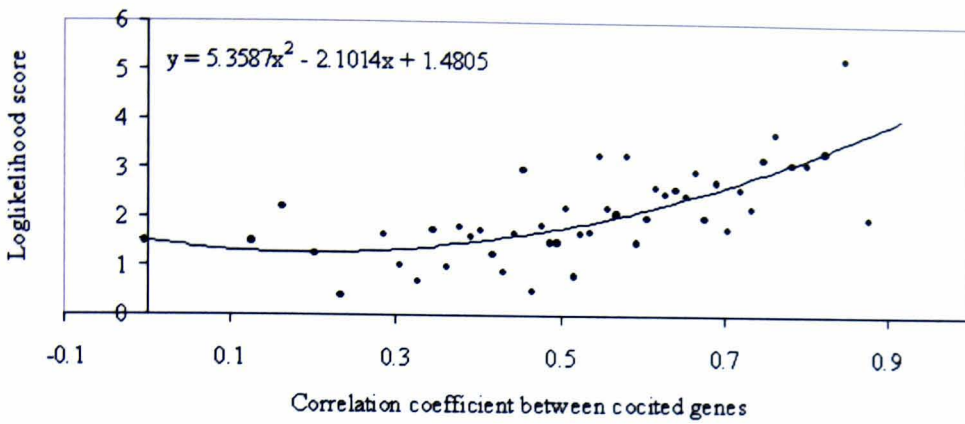


Figure D.1: Log likelihood score calculated for cocited genes based on KEGG pathway benchmark using bins of size 215 pairs.

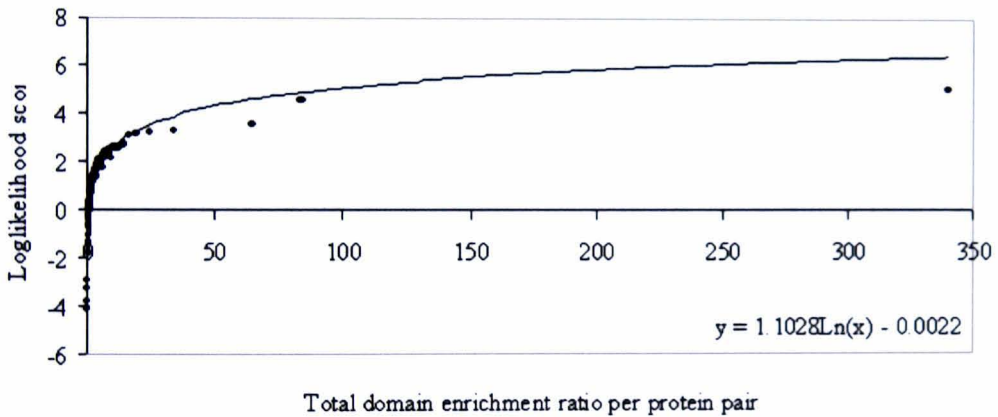


Figure D.2: Log likelihood score calculated for proteins sharing common interpro domains, based on KEGG pathway benchmark using bins of size 4377 pairs. A protein pair is based on whether two proteins share one or more domains. The domain enrichment ratio per shared domain is summed to calculate the final weight for the pair.

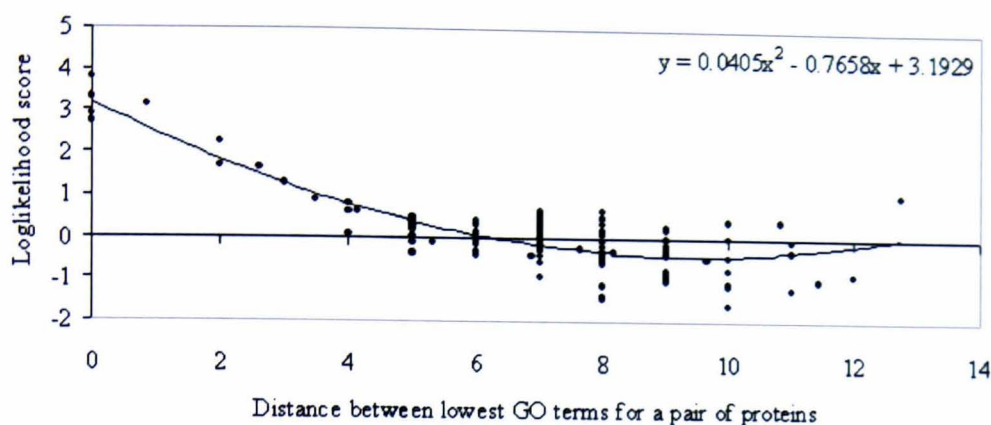


Figure D.3: Log likelihood score based on the distance between the lowest GO terms for a pair of proteins compared to KEGG pathway benchmark. A minimum GO term depth of level 7 was considered. Bins of size 8899 pairs were used.

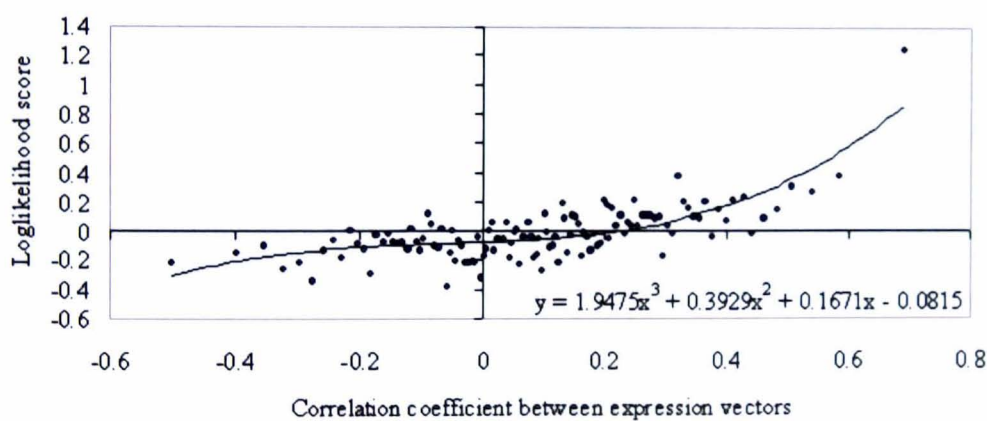


Figure D.4: Log likelihood score calculated for co-expressed genes based on KEGG pathway benchmark using bins of size 69343 pairs.

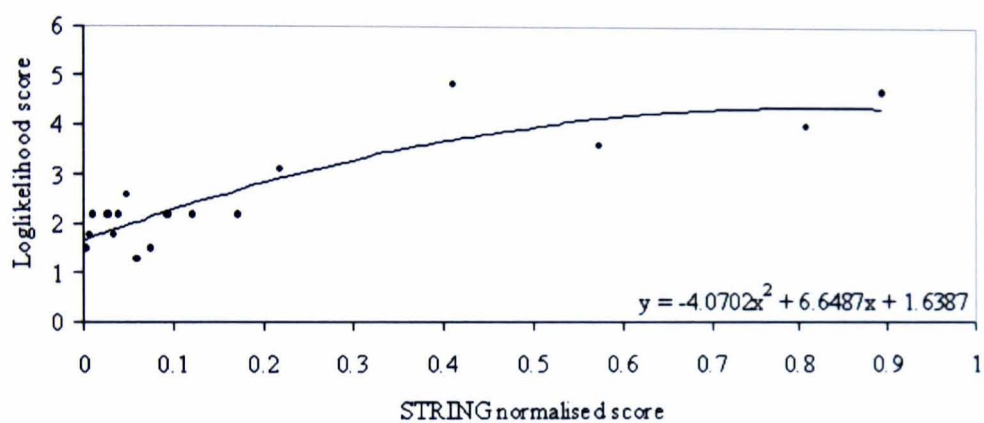


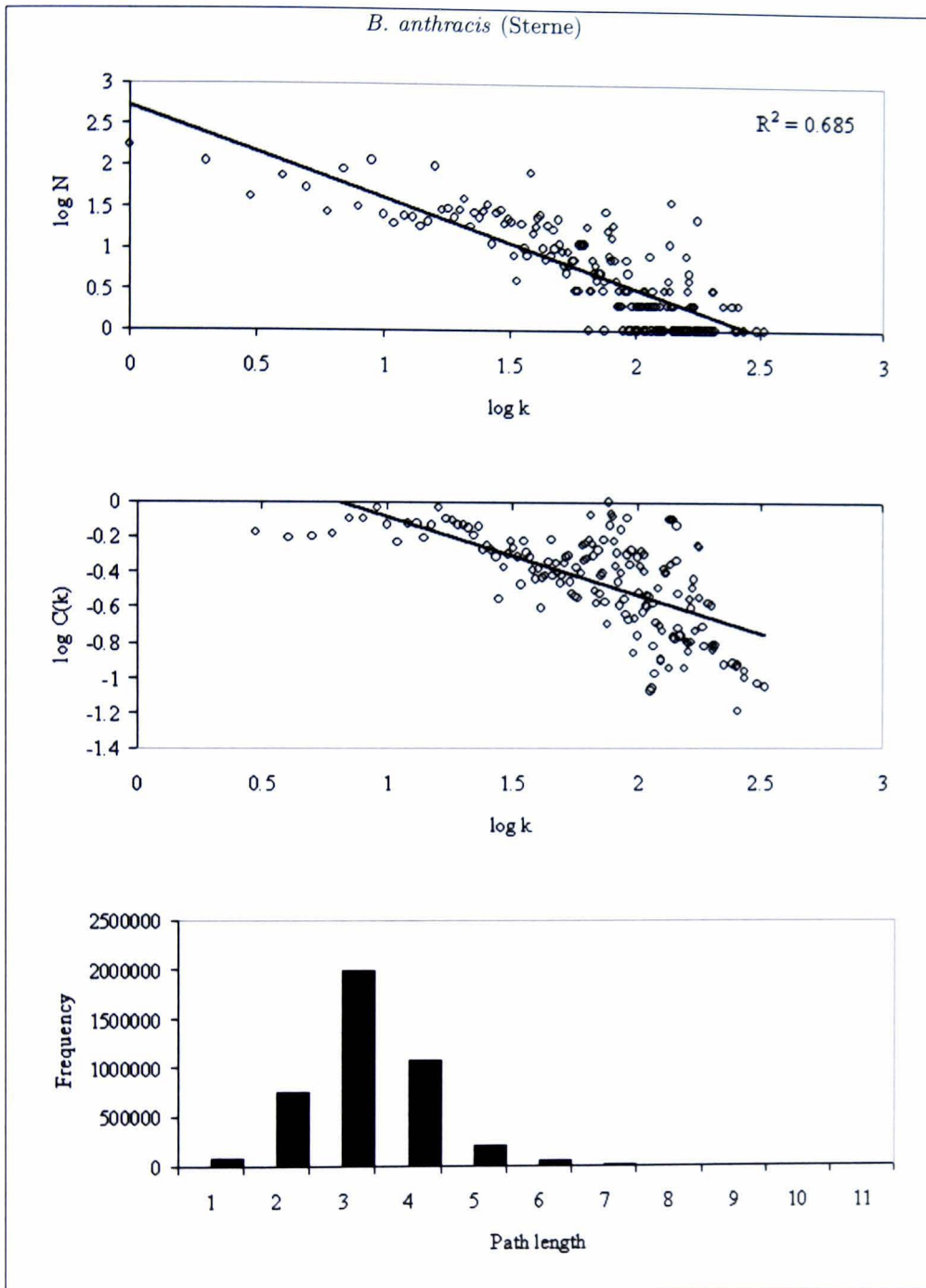
Figure D.5: Log likelihood score for the normalised STRING fusion data based on KEGG pathway benchmark using bins of size 47 pairs.

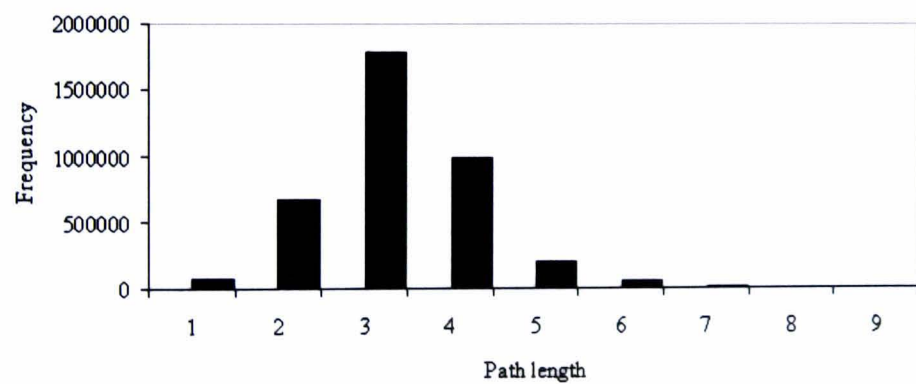
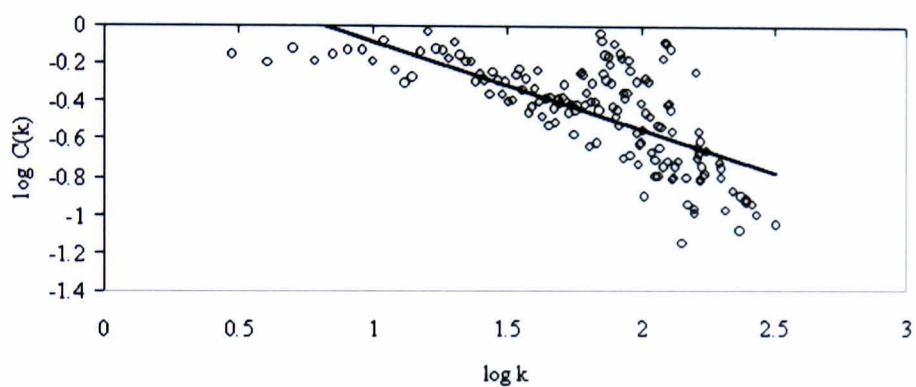
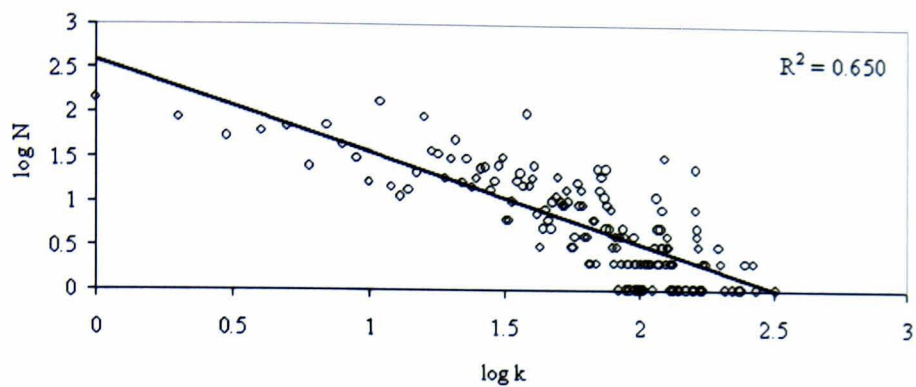
Appendix E

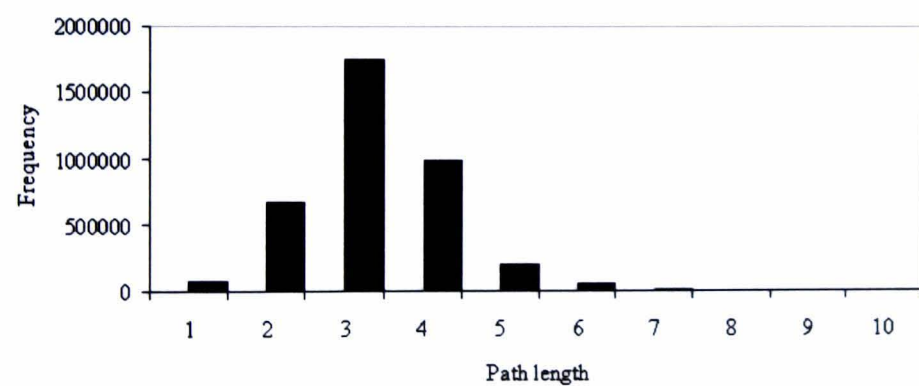
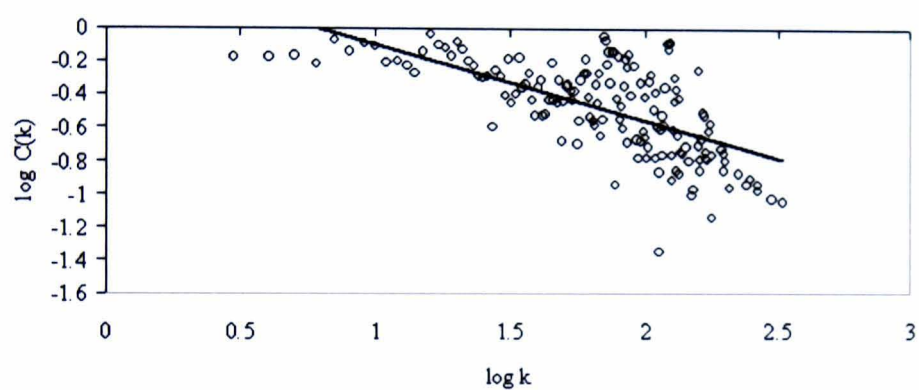
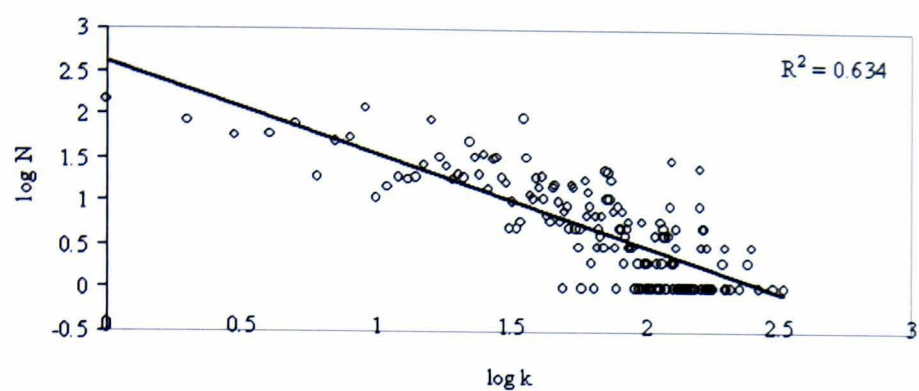
SubtilNet network topologies

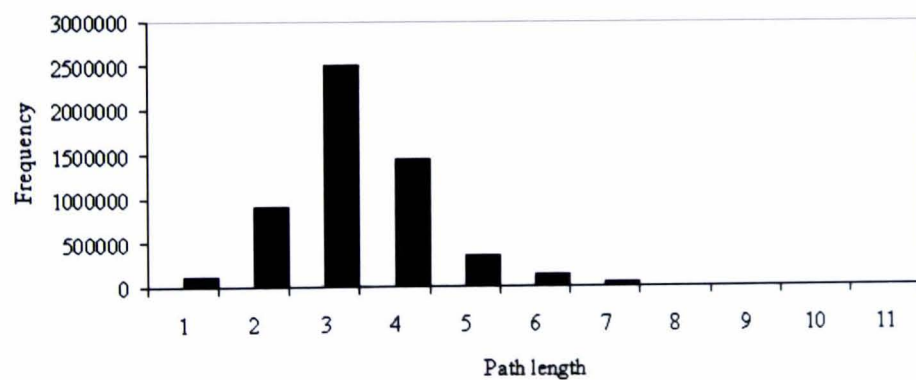
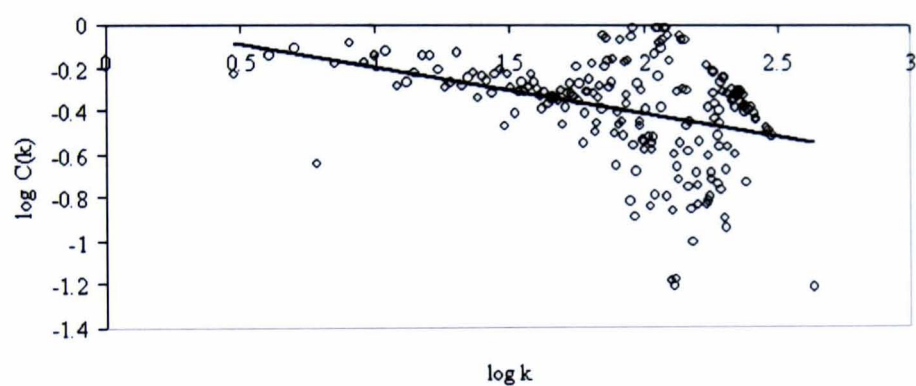
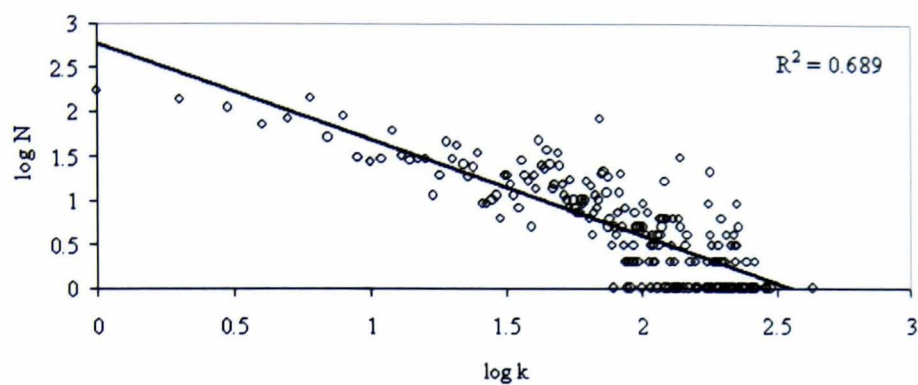
Table E.1 shows four distribution plots describing the network topology for each *Bacillus* PFIN. Top left graph shows the node degree distribution, top right shows the average clustering coefficient distribution, bottom left shows the shortest path length distribution.

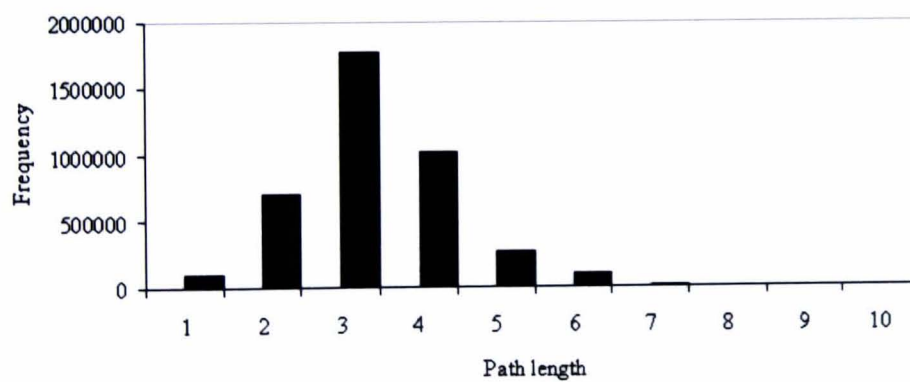
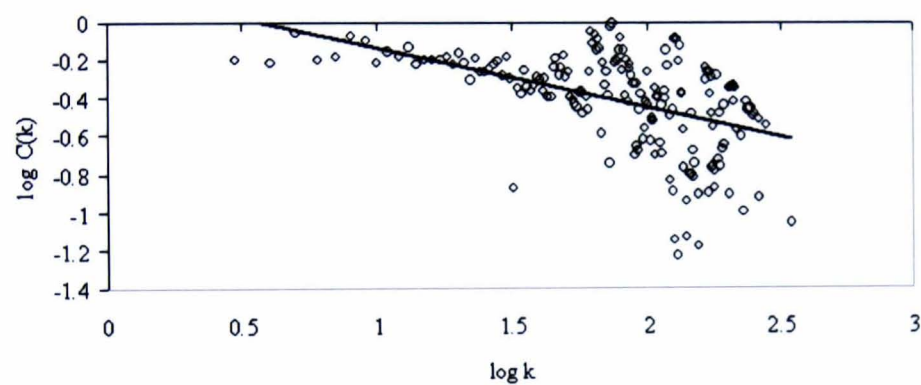
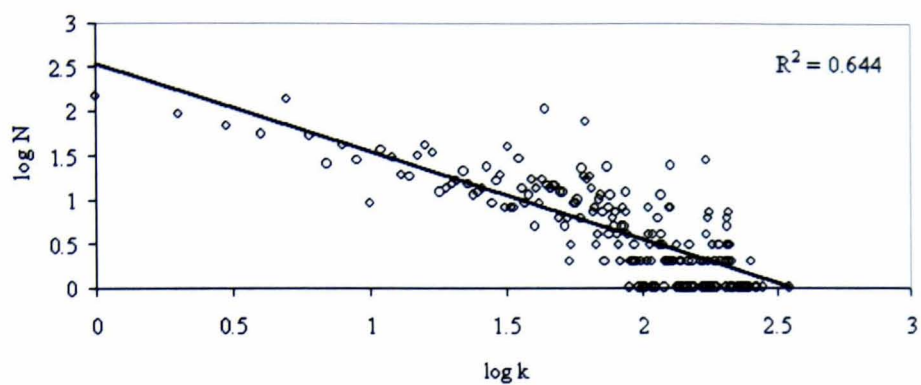
Table E.1: Network topology.

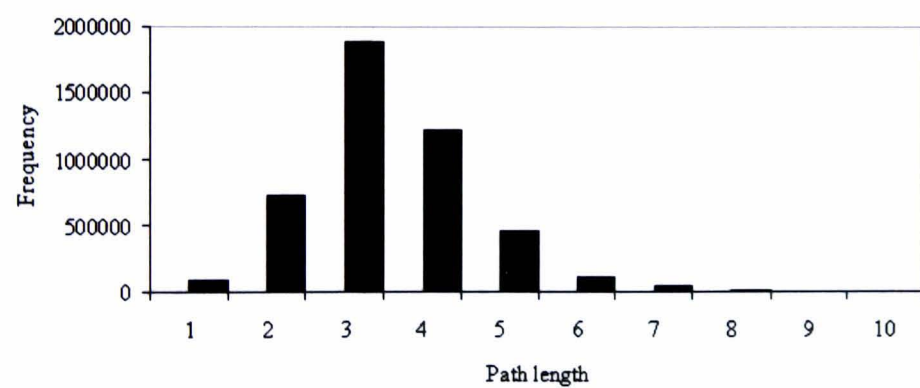
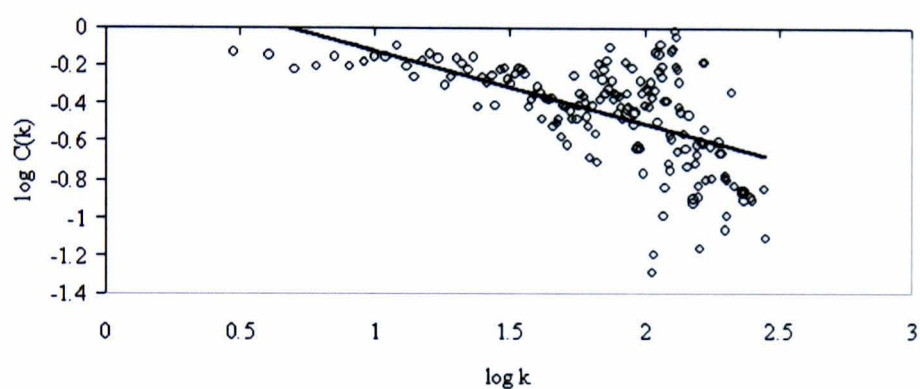
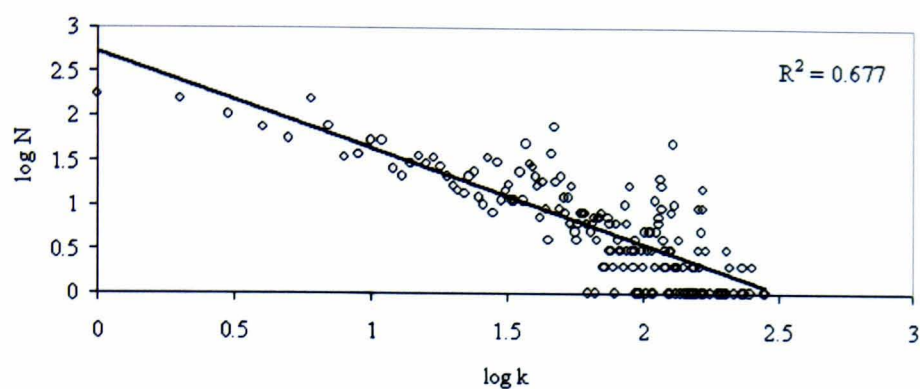


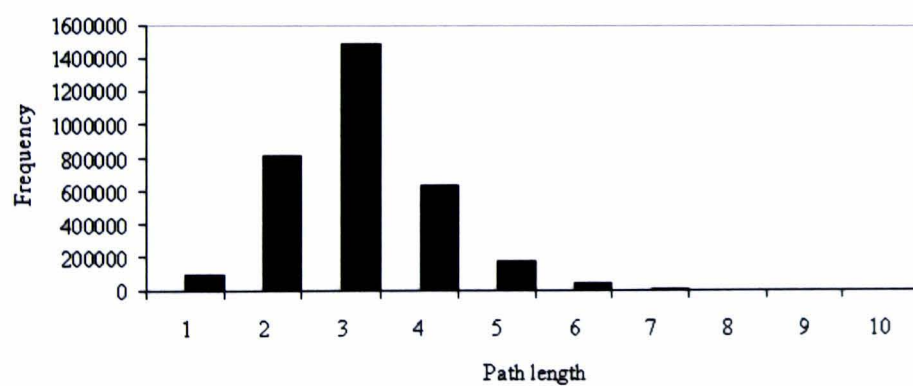
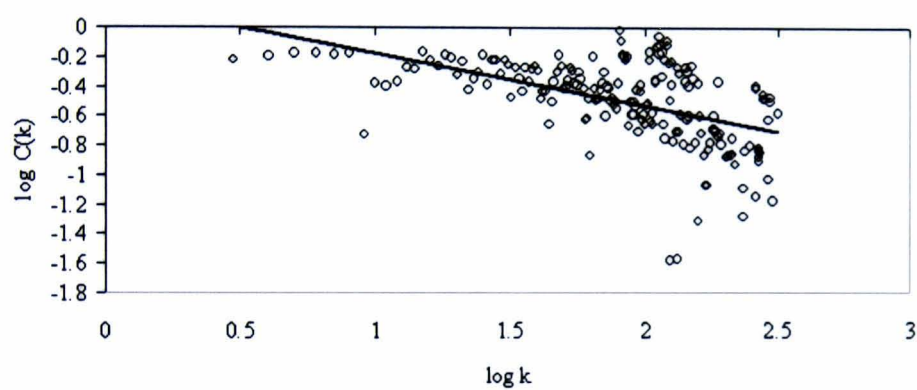
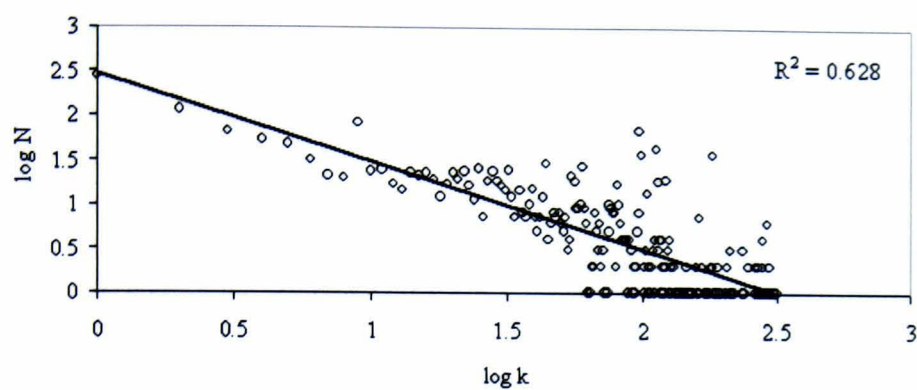
B. anthracis (Ames ancestor)

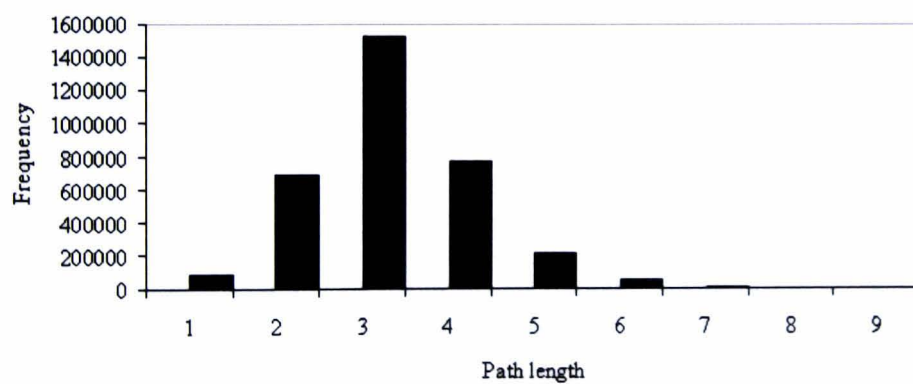
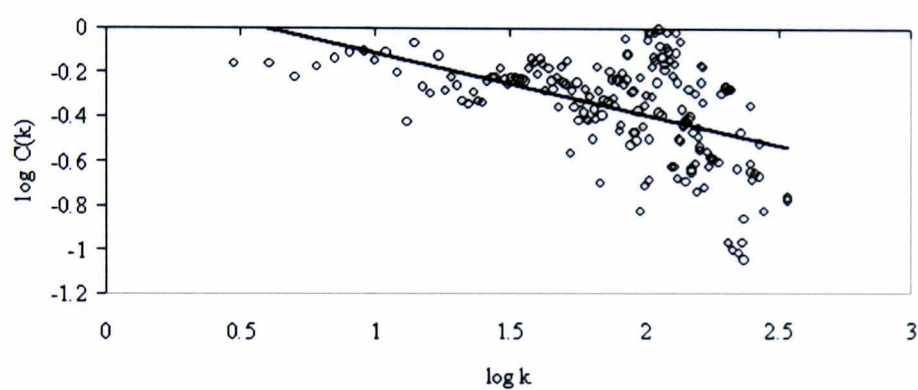
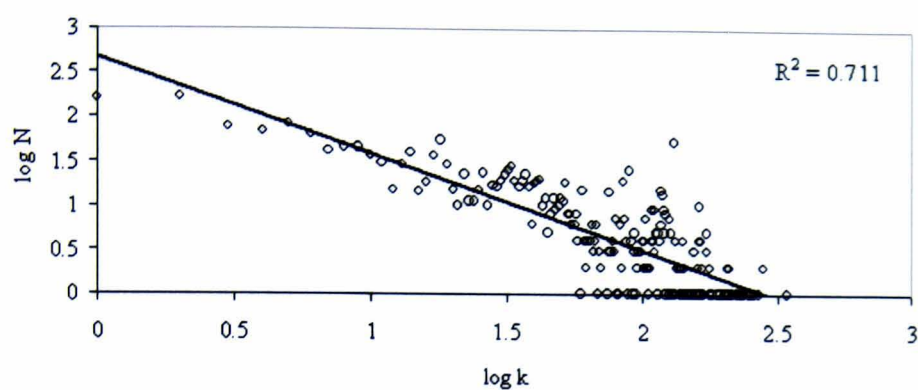
B. anthracis (Ames)

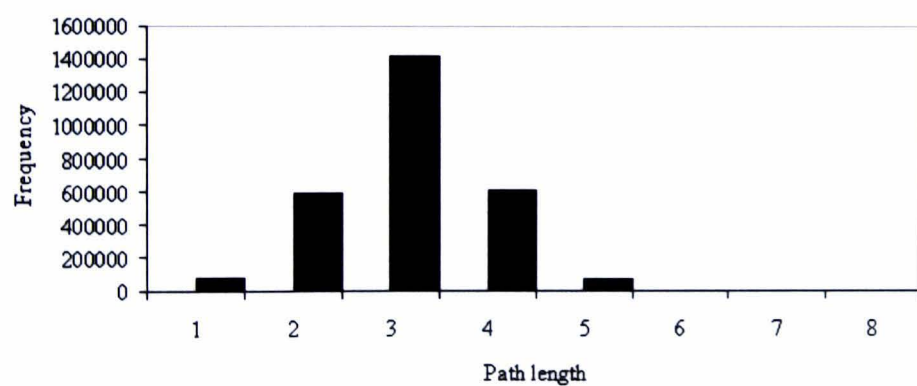
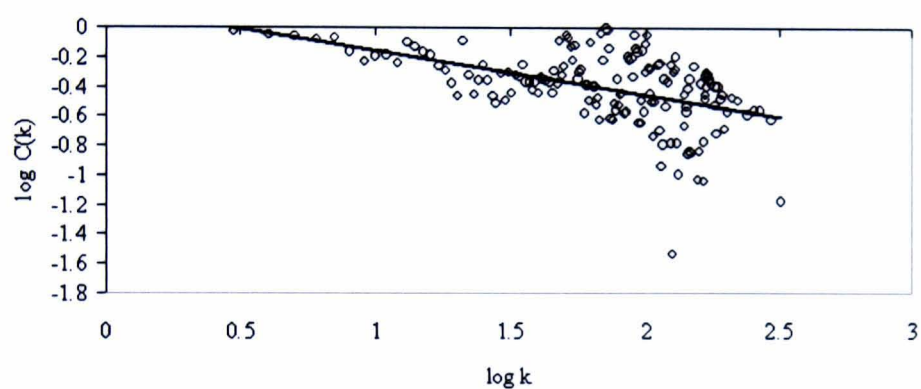
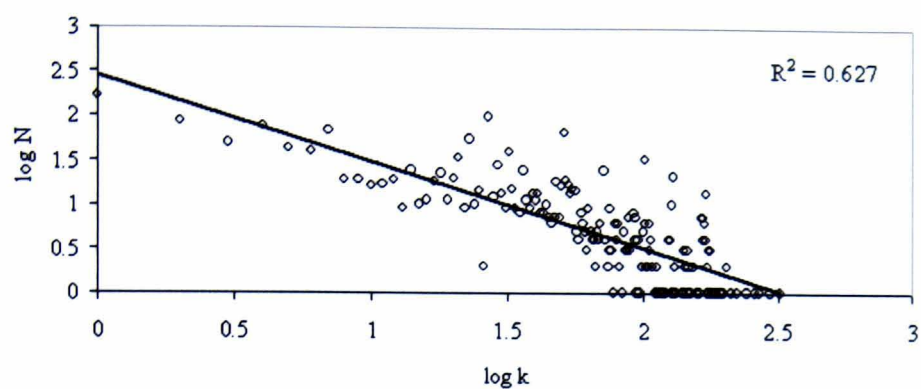
B. cereus (E33L)

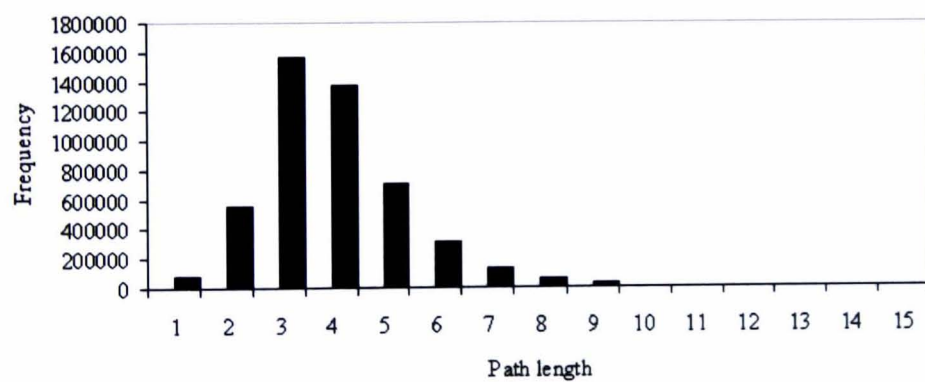
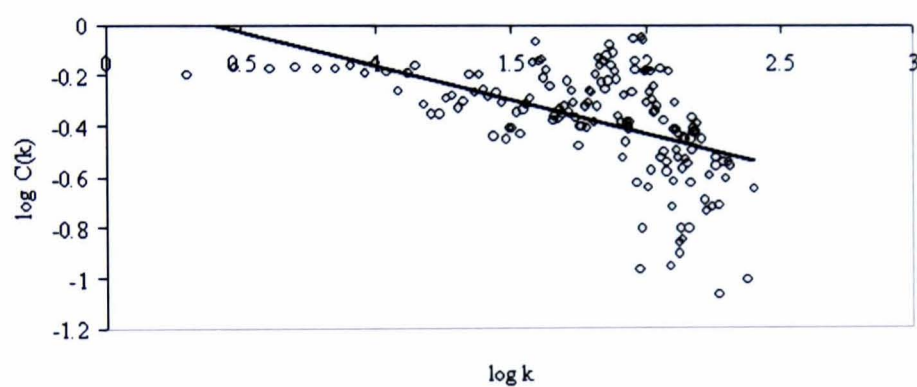
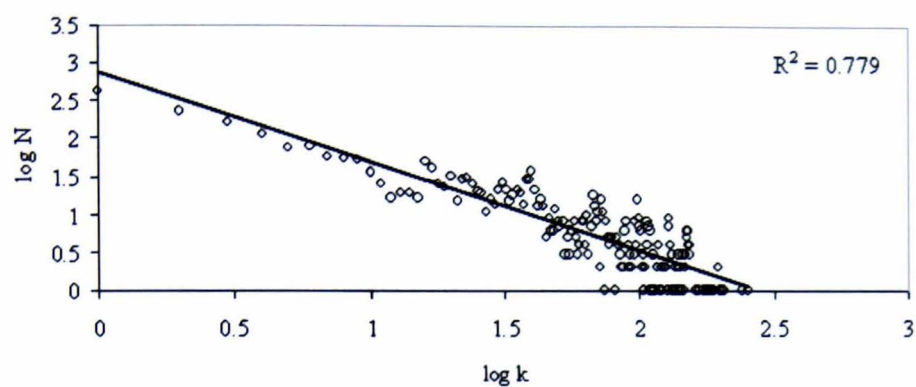
B. cereus (ATCC 10987)

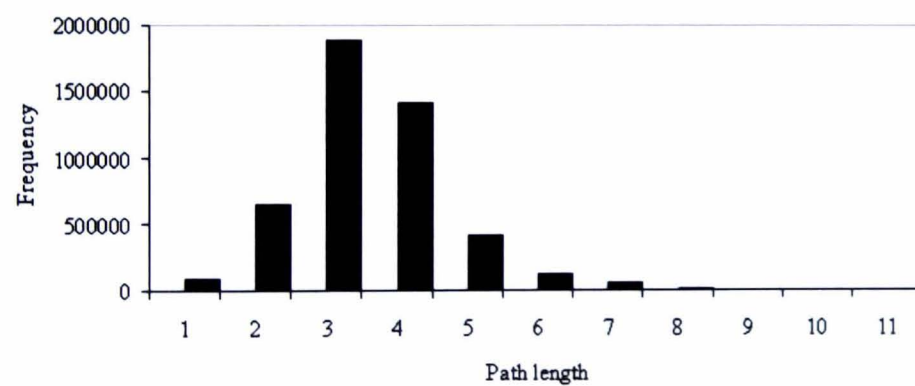
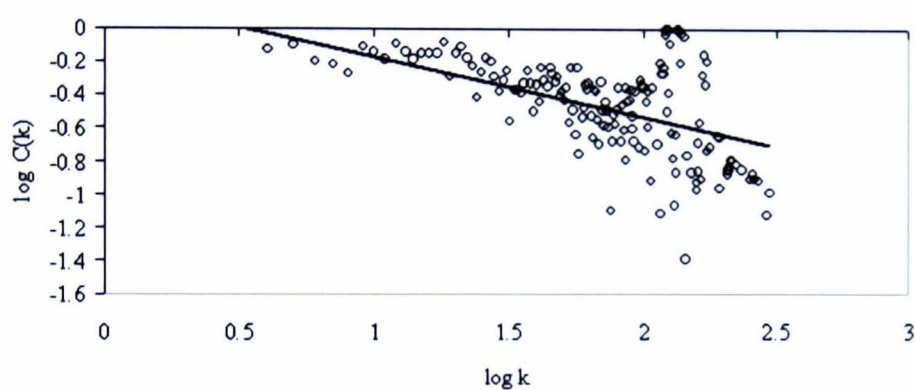
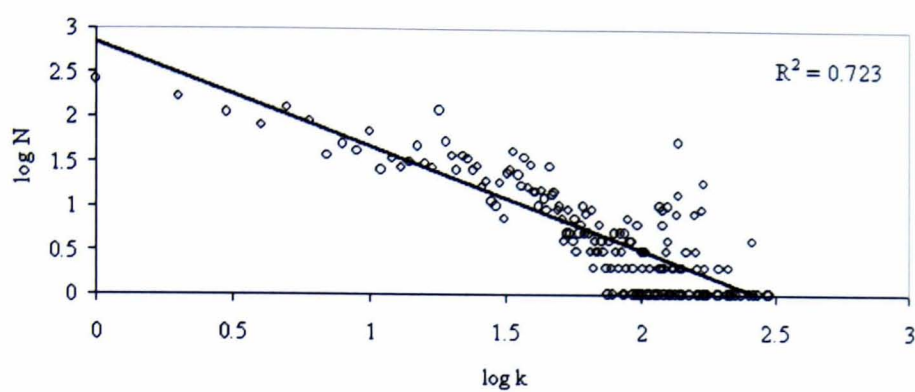
B. cereus (ATCC 14579)

B. clausii (KSM-K16)

B. halodurans (C-125)

B. licheniformis (ATCC 14580)

B. subtilis (strain 168)

B. thuringiensis konkukian (strain 97-27)

Appendix F

B. subtilis PFIN clusters

A more detailed breakdown of the clusters obtained using the MCODE plugin for Cytoscape. This table highlights the different functional categories represented in each cluster and their proportions, as well as indicating those proteins predicted to be secreted by BaSPP.

Table F.1: Detailed view of the major clusters found in the *B. subtilis* PFIN. Gene names in bold highlights those whose products are predicted to be secreted; esample descriptions are given for these genes.

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
1	99	1.0	7	Sporulation, germination and cell division	1.0	<i>spoIIAB</i>
				Adaptation to atypical conditions	3.0	<i>rsbW htpG rsbT</i>
				DNA replication, restriction/modification and repair	1.0	<i>mutL</i>
				Mobility and chemotaxis	22.2	<i>motB tlpB fliM cheV motA tlpC cheW mcpB fliG cheD yvaQ yfmS mcpA cheR ytzE cheC mcpC yoaH ytzD hemAT fliY tlpA</i>
				DNA recombination, packaging and segregation	2.0	<i>parE gyrB</i>
				Sensors (signal transduction)	36.3	<i>kinA degS phoR dctS yrkQ lytS ykoH kinE ytsB resE yocF yxjM yccG comP ycbA yvqE yxdK yvfT cssS ydfH yycG kinC citS kinD ywpD yufL yvcQ yfiJ cheA yhcY yclK ycbM ybdK yvrG kinB yesM</i>
				RNA synthesis and modification	33.3	<i>yocG yccH yxjL yrkP yneI ybdJ lytT cssR yxdJ yycF dctR ytsA resD citT yhcZ ycbL yvrH spo0A yclJ cheB ycbB ykoG yvfU yvcP ydfI yfiK degU yufM spo0F yesN comA phoP yvqC</i>
				Unassigned	1.0	<i>ykoK</i>

2	59	3.3	6	Cell wall	3.3	<i>dltC dltA</i>
				Mobility and chemotaxis	1.6	<i>yvzB</i>
				Unknown	5.0	<i>yvmC yrdC yddQ</i>
				Unassigned	1.6	<i>yccK</i>
				Antibiotic production	23.7	<i>ppsA srfAC ppsC acpK pksM</i> <i>ppsB ppsE pksN pksL ppsD</i> <i>srfAB srfAA pksR pksJ</i>
				Metabolism of carbohydrates and related molecules	32.2	<i>ykvO yxjF yvgN gdh yrpG</i> <i>yqkF yhdF ytbE yhxC ycsN</i> <i>kduD yxbG yxnA ycdF yhxD</i> <i>yusZ yuxG ykuF yjmF</i>
				Metabolism of coenzymes and prosthetic groups	11.8	<i>dhbA yueJ dhbB yaaI dhbE</i> <i>ywoC dhbF</i>
3	41	14.6	5	Metabolism of lipids	20.3	<i>yvrD ymfI ytkK ygjQ yvaG</i> <i>yusS yoxD yjdA fabL acpA</i> <i>fabG ywfH</i>
				Adaptation to atypical conditions	2.4	<i>yveN</i>
				Cell wall	4.8	<i>yngB gtaB</i>
				Transport/binding proteins and lipoproteins	7.3	<i>ybbF ptsG sacP</i>
				Unknown	4.8	<i>yvdK yoxA</i>
				Unassigned	19.5	<i>treP malP nagP gamP yvgW</i> <i>yvgX yloB ykvW</i>
				Protein synthesis and modification	2.4	<i>hprP</i>
2	59	3.3	6	Metabolism of carbohydrates and related molecules	58.5	<i>glgD xylA yckE yvdF treA</i> <i>bglA yveB bglH amyE sacC</i> <i>sacX bglC amyX ydhP sacA</i> <i>glgB yugT malL yhcW ycdG</i> <i>yveP yfnH sacB pgcM</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
4	30	0.0	4	DNA replication, restriction/modification and repair	6.6	<i>yorL yshC</i>
				Metabolism of nucleotides and nucleic acids	16.6	<i>adk purK purF upp purC</i>
				Protein synthesis and modification	73.3	<i>rplV rplD rpsH rpsC rpsJ rplN rplB rplP map rpsE rplX rplO rpmC rplW rpsN rplE rplR rpsQ rplF rplC rpmD rpsS</i>
				Metabolism of coenzymes and prosthetic groups	3.3	<i>pabA</i>
5	40	0.0	7	Transformation/competence	2.5	<i>comEB</i>
				Sporulation, germination and cell division	2.5	<i>splB</i>
				Cell wall	7.5	<i>murE murF murC</i>
				Metabolism of amino acids and related molecules	40.0	<i>hisC thrB yhaA ykuQ dapG kbl hisH yxeP lysC dapF yqjE ykuR dapA yclM dapB pepT</i>
				Unknown	10.0	<i>yvdG ywnB yfkA yunD</i>
				Metabolism of nucleotides and nucleic acids	30.0	<i>ytnL purT purD cdd pyrC cmk udk pdp hprT pyrG yfkN yhcR</i>
				Protein synthesis and modification	5.0	<i>amhX ytiP</i>
				Metabolism of coenzymes and prosthetic groups	2.5	<i>ykuL</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
6	81	1.2	11	Transformation/competence	1.2	<i>comGA</i>
				DNA replication, restriction/modification and repair	2.4	<i>radA dnaA</i>
				Cell wall	1.2	<i>tagH</i>
				Membrane bioenergetics (electron transport chain and ATP synthase)	7.4	<i>yojN atpH atpF qoxC atpD qoxB</i>
				Unknown	6.1	<i>yruN ykqA ykuS yfnB ysaA</i>
				Phage-related functions	2.4	<i>xkdC yqaM</i>
				Adaptation to atypical conditions	4.9	<i>clpC clpY clpE clpX</i>
				Sporulation, germination and cell division	7.4	<i>ytpT spoVK spoIIIE ftsE spoIIIAA cotA</i>
				DNA recombination, packaging and segregation	2.4	<i>yrrC ruvB</i>
				Transport/binding proteins and lipoproteins	6.1	<i>glnQ fhuC expZ rbsA yfiL</i>
				RNA synthesis and modification	3.7	<i>levR gutR rocR</i>
				Unassigned	53.0	<i>yvcR yybJ mntB ycdI ydbJ yhaQ yqiZ yknY yclH yjkB yydI ywjA yzlF yurJ ykpA ytmN yfmF yufO ytsC ylmA yusC ecsA ytlC yckI natA ythP ssuB yvrA yfmM yusV ygaD yclP yvcC ybdA ydiF msmX yvrO ykoD yvfR sunT yzeO yzdL yfmR</i>
				Metabolism of carbohydrates and related molecules	1.2	<i>sdhC</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
7	65	3.0	10	DNA replication, restriction/modification and repair	1.5	<i>polC</i>
				Cell wall	6.1	<i>murD murB murAA murAB</i>
				Detoxification	6.1	<i>yisY yaaD yrhJ yetO</i>
				Transport/binding proteins and lipoproteins	3.0	<i>glnP yyzE</i>
				Metabolism of amino acids and related molecules	23.0	<i>carA hisB aroK aroE aspB pheA aroB tyrA yqhS trpA trpF trpD aroD trpC ureC</i>
				RNA synthesis and modification	1.5	<i>ctsR</i>
				Unassigned	1.5	<i>ypqE</i>
				Metabolism of nucleotides and nucleic acids	9.2	<i>dck apt ypfD pyrAA yjbP purQ</i>
				Protein synthesis and modification	21.5	<i>hisZ glyS fnt tyrS efp leuS metS gltX thrS pheT ytzM valS pheS trpS</i>
				Metabolism of carbohydrates and related molecules	18.4	<i>csn gamA nagB nagA yvaM yfhM ybbT ydhS pmi manA yvyH araM</i>
				Metabolism of coenzymes and prosthetic groups	7.6	<i>bioA hemA bioB bioD bioW</i>
8	11	18.1	2	Metabolism of carbohydrates and related molecules	9.0	<i>gpsA</i>
				Metabolism of lipids	90.9	<i>pgsA ywiE psd yhdW dgkA cdsA ywnE yqiK pssA glpQ</i>
9	17	5.8	4	Sporulation, germination and cell division	17.6	<i>ydhD cotSA yaaH</i>
				Cell wall	29.4	<i>ypjH tagE tuaC tuaH ytcC</i>
				Unknown	17.6	<i>ycsE yqgM yvbX</i>
				Metabolism of carbohydrates and related molecules	29.4	<i>glgA ywqF lacA yesZ abfA</i>
				Metabolism of lipids	5.8	<i>ugtP</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
10	11	0.0	6	Cell wall	9.0	<i>yqfY</i>
				Detoxification	9.0	<i>yqfP</i>
				Unknown	18.1	<i>yacM yacN</i>
				RNA synthesis and modification	9.0	<i>comQ</i>
				Metabolism of carbohydrates and related molecules	18.1	<i>dxs dxr</i>
				Metabolism of coenzymes and prosthetic groups	9.0	<i>hepT</i>
				Metabolism of lipids	27.2	<i>yisP uppS yqiD</i>
11	43	2.3	9	Sporulation, germination and cell division	6.9	<i>yjoB ftsH spoIIIE</i>
				Adaptation to atypical conditions	4.6	<i>lonA lonB</i>
				DNA replication, restriction/modification and repair	6.9	<i>dnaC dnaI uvrA</i>
				Mobility and chemotaxis	2.3	<i>fliI</i>
				DNA recombination, packaging and segregation	2.3	<i>recA</i>
				Membrane bioenergetics (electron transport chain and ATP synthase)	4.6	<i>ndhF cccA</i>
				Transport/binding proteins and lipoproteins	4.6	<i>dppD ybaE</i>
				Unknown	2.3	<i>yhaN</i>
				RNA synthesis and modification	9.3	<i>yplP rho acoR bkdR</i>
				Unassigned	51.1	<i>appF pstBB pstBA opuAA ykfD appD oppD cydD cydC yhcG yhcH yheH yheI opuBA yfiC yfiB yknU yknV ytrE ytrB ybzA opuCA</i>
				Protein secretion	4.6	<i>ftsY ffh</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
12	27	7.4	9	Adaptation to atypical conditions	3.7	<i>rsbQ</i>
				Detoxification	7.4	<i>nap ybfK</i>
				Metabolism of amino acids and related molecules	3.7	<i>ybaC</i>
				Metabolism of nucleotides and nucleic acids	3.7	<i>deoD</i>
				Antibiotic production	7.4	<i>pksE pksC</i>
				Protein synthesis and modification	3.7	<i>asnS</i>
				Metabolism of carbohydrates and related molecules	18.5	<i>ycbF ycgS ykfB yitF yqiD</i>
				Metabolism of coenzymes and prosthetic groups	29.6	<i>menC hemH nasF ylnD hemE hemY ylnF menA</i>
				Metabolism of lipids	22.2	<i>yclB fabF accD ytpA accB yusL</i>
13	17	11.7	7	DNA replication, restriction/modification and repair	17.6	<i>ydiO ydiP mtbP</i>
				Metabolism of sulfur	5.8	<i>yvgR</i>
				Sensors (signal transduction)	5.8	<i>luxS</i>
				Unknown	11.7	<i>ykrX ykrZ</i>
				Metabolism of amino acids and related molecules	35.2	<i>ytkP ykrV yrhA cysE serA hom</i>
				Metabolism of nucleotides and nucleic acids	5.8	<i>mtn</i>
				Protein synthesis and modification	5.8	<i>ykrS</i>
				Metabolism of carbohydrates and related molecules	11.7	<i>ykrW yoaD</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
14	22	18.1	7	Metabolism of phosphate	4.5	<i>phoD</i>
				Cell wall	4.5	<i>yqiI</i>
				Detoxification	4.5	<i>mrpD</i>
				Transport/binding proteins and lipoproteins	4.5	<i>yvrP</i>
				Membrane bioenergetics (electron transport chain and ATP synthase)	27.2	<i>atpB qcrA atpE yhfW ctaA atpC</i>
				Unknown	9.0	<i>yvaX ywbN</i>
				Metabolism of nucleotides and nucleic acids	4.5	<i>pyrE</i>
				Metabolism of carbohydrates and related molecules	40.9	<i>abnA ytcB rpe ydjE xylB gntK tkt araD araL</i>
15	8	12.5	3	Adaptation to atypical conditions	12.5	<i>yveR</i>
				RNA synthesis and modification	37.5	<i>licR mtlR manR</i>
				Metabolism of carbohydrates and related molecules	50.0	<i>yveO ydhT yveT yulE</i>
16	8	12.5	1	Mobility and chemotaxis	100.0	<i>fliP fliL flhA flhB fliZ ylxH flhF fliR</i>
17	19	5.2	4	Membrane bioenergetics (electron transport chain and ATP synthase)	36.8	<i>ctaE qcrC ctaF ctaD qcrB ctaO ctaB</i>
				Metabolism of amino acids and related molecules	5.2	<i>kamA</i>
				Unknown	21.0	<i>ymcB yloN ytqA yqeV</i>
				Antibiotic production	5.2	<i>albA</i>
				Metabolism of coenzymes and prosthetic groups	31.5	<i>hemN ybcP moaA lipA bioF hemZ</i>
18	13	15.3	3	Cell wall	30.7	<i>lytC mraY yrvJ tagO</i>
				Detoxification	7.6	<i>yubB</i>
				Unknown	15.3	<i>ywjB yyaP</i>
				Metabolism of coenzymes and prosthetic groups	46.1	<i>dfrA folB ykuK sul folK folC</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
19	11	18.1	4	Detoxification	9.0	<i>ykfA</i>
				Transport/binding proteins and lipoproteins	45.4	<i>appC oppA appA appB dppC</i>
				Unassigned	27.2	<i>oppB oppC oppF</i>
				Metabolism of carbohydrates and related molecules	9.0	<i>ykfC</i>
				Protein synthesis and modification	9.0	<i>dppA</i>
20	7	0.0	1	Metabolism of carbohydrates and related molecules	100.0	<i>pgm fbaA pgi eno pyk pgk gapA</i>
21	21	9.5	5	Metabolism of phosphate	9.5	<i>phoB phoA</i>
				RNA synthesis and modification	4.7	<i>pyrR</i>
				Unknown	4.7	<i>yorR</i>
				Metabolism of amino acids and related molecules	19.0	<i>rocD ansB argB carB</i>
				Metabolism of nucleotides and nucleic acids	52.3	<i>purL guaB thyB pyrD tmk thyA yncF yosS guaA pyrF pyrAB</i>
				Protein synthesis and modification	9.5	<i>lysS aspS</i>
22	30	6.6	9	Sporulation, germination and cell division	6.6	<i>spnL spoVFB</i>
				Detoxification	16.6	<i>cypX pksS yjiB cypA ydjP</i>
				RNA synthesis and modification	20.0	<i>ydqJ ywhA ysmB ykvE ykoM yfiV</i>
				Metabolism of amino acids and related molecules	10.0	<i>yurR yclE goxB</i>
				Unknown	13.3	<i>ykmA yraK yqjL ybfA</i>
				Metabolism of carbohydrates and related molecules	10.0	<i>xsa yomI glpD</i>
				Metabolism of coenzymes and prosthetic groups	6.6	<i>bioI yueK</i>
				Phage-related functions	10.0	<i>yqbO xkdO yjbJ</i>
				Metabolism of lipids	3.3	<i>ywjE</i>
				Transposon and IS	3.3	<i>yddH</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
23	12	0.0	3	Cell wall	8.3	<i>ddl</i>
				Metabolism of amino acids and related molecules	33.3	<i>yncD alr dsdA thrC</i>
				Protein synthesis and modification	58.3	<i>proS alaS hisS ytpR cysS serS ileS</i>
24	14	0.0	4	Unknown	21.4	<i>yqeJ ytaG yoaP</i>
				Metabolism of amino acids and related molecules	35.7	<i>ylmB yodQ ycgN rocA ilvC</i>
				Metabolism of carbohydrates and related molecules	14.2	<i>citG icd</i>
				Metabolism of coenzymes and prosthetic groups	21.4	<i>panC coaA panB</i>
				Metabolism of lipids	7.1	<i>acpS</i>
25	11	0.0	5	Metabolism of amino acids and related molecules	27.2	<i>aroF pheB trpE</i>
				Antibiotic production	9.0	<i>pksD</i>
				Metabolism of carbohydrates and related molecules	9.0	<i>pycA</i>
				Metabolism of coenzymes and prosthetic groups	45.4	<i>yueD dhhC menF menE pabB</i>
				Metabolism of lipids	9.0	<i>lcfA</i>
26	8	50.0	4	Adaptation to atypical conditions	12.5	<i>ywsC</i>
				Sporulation, germination and cell division	12.5	<i>spsG</i>
				Detoxification	12.5	<i>yjiC</i>
				Cell wall	37.5	<i>cwlD lytD murG</i>
				Unknown	25.0	<i>yraJ yraI</i>
27	8	0.0	4	Detoxification	25.0	<i>katX katE</i>
				RNA synthesis and modification	12.5	<i>tenI</i>
				Unknown	25.0	<i>yjbN ytdI</i>
				Metabolism of coenzymes and prosthetic groups	25.0	<i>nadA nadC</i>
				Metabolism of lipids	12.5	<i>des</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
28	6	0.0	1	Metabolism of coenzymes and prosthetic groups	100.0	<i>thiE thiC yjbV thiD thiL thiM</i>
29	6	0.0	1	Metabolism of nucleotides and nucleic acids	100.0	<i>pucH pucB pucD pucC pucM pucE</i>
30	6	0.0	2	Adaptation to atypical conditions	83.3	<i>rsbV rsbX rsbR rsbS rsbU</i>
				RNA synthesis and modification	16.6	<i>sigB</i>
31	6	16.6	1	Mobility and chemotaxis	83.3	<i>fliJ ylxG fliH flgE fliK</i>
				Unknown	16.6	<i>ylxF</i>
32	6	50.0	1	Transformation/competence	100.0	<i>comGD comGC comGB comGF comGG comGE</i>
33	6	83.3	1	Transport/binding proteins and lipoproteins	16.6	<i>fhuD</i>
				Unassigned	83.3	<i>yfiY yclQ yzeB yfmC yhfQ</i>
34	6	0.0	2	Metabolism of amino acids and related molecules	16.6	<i>yobN</i>
				Unknown	33.3	<i>yabC yjjA</i>
				Metabolism of coenzymes and prosthetic groups	50.0	<i>hemD gsaB hemL</i>
35	7	0.0	1	Metabolism of amino acids and related molecules	100.0	<i>proJ argD proA argJ hutI rocF argF</i>
36	13	38.4	2	Unknown	46.1	<i>yorA ywoF yfiA yclG yomE ycbH</i>
				Metabolism of carbohydrates and related molecules	46.1	<i>ycbC pel yvpA pelB hzlA uzaC</i>
				Phage-related functions	7.6	<i>yobO</i>
37	5	0.0	0	Unassigned	100.0	<i>fruA ydhO ywbA licC manP</i>
38	5	0.0	1	Metabolism of nucleotides and nucleic acids	100.0	<i>purN purH purM gmk guaC</i>
39	5	20.0	1	Unassigned	60.0	<i>araP araQ araN</i>
				Metabolism of carbohydrates and related molecules	40.0	<i>araA araB</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
40	5	0.0	1	Transport/binding proteins and lipoproteins	20.0	<i>opuCD</i>
				Unassigned	80.0	<i>opuAB opuBD opuCB opuBB</i>
41	6	33.3	0	Unassigned	100.0	<i>yqiX yckA yckJ yxeN yckK ytmM</i>
42	5	0.0	1	Unknown	80.0	<i>yzlG yzlD yzlC yzlE</i>
				RNA synthesis and modification	20.0	<i>sigY</i>
43	5	20.0	1	Antibiotic production	100.0	<i>albF albG albD albE albB</i>
44	5	20.0	2	Detoxification	80.0	<i>mrpE mrpF mrpG mrpB</i>
				Transport/binding proteins and lipoproteins	20.0	<i>mrpA</i>
45	5	20.0	2	Sporulation, germination and cell division	40.0	<i>minD minC</i>
				Cell wall	60.0	<i>mreB mreC mreD</i>
46	5	0.0	1	Metabolism of coenzymes and prosthetic groups	100.0	<i>moaD mobA mobB moaE moeA</i>
47	5	0.0	1	Adaptation to atypical conditions	100.0	<i>ykoB yqhA yojH yetI yezB</i>
48	5	20.0	1	RNA synthesis and modification	20.0	<i>yesS</i>
				Unknown	40.0	<i>yesU yesW</i>
				Unassigned	40.0	<i>yesO yesQ</i>
49	5	0.0	1	Sporulation, germination and cell division	100.0	<i>cotV cotX cotY cotZ cotW</i>
50	5	60.0	1	Metabolism of amino acids and related molecules	100.0	<i>ispA aprX aprE bpr vpr</i>
51	5	0.0	2	Adaptation to atypical conditions	40.0	<i>ydaG yocK</i>
				Unknown	20.0	<i>yfkM</i>
				Metabolism of carbohydrates and related molecules	40.0	<i>yhdN ydaD</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
52	12	0.0	3	Unknown	41.6	<i>yteT yvaA yhjJ yulF yfiI</i>
				Metabolism of amino acids and related molecules	8.3	<i>yrbE</i>
				Metabolism of carbohydrates and related molecules	8.3	<i>yktC</i>
				Metabolism of lipids	41.6	<i>lpdV bkdAB bcd bkdAA ptb</i>
53	5	20.0	2	Membrane bioenergetics (electron transport chain and ATP synthase)	80.0	<i>ctaC ythA cccB cydA</i>
				Metabolism of amino acids and related molecules	20.0	<i>nasE</i>
54	5	0.0	2	Metabolism of carbohydrates and related molecules	80.0	<i>acoL citZ citA pdhD</i>
				Metabolism of lipids	20.0	<i>bkdB</i>
55	5	0.0	2	RNA synthesis and modification	20.0	<i>rpoA</i>
				Protein synthesis and modification	80.0	<i>rpsK rpmJ rplQ infA</i>
56	5	0.0	1	Mobility and chemotaxis	100.0	<i>hag fliT fliD yvyC fliS</i>
57	5	0.0	1	Sporulation, germination and cell division	100.0	<i>sspF sspD sspC sspB sspA</i>
58	8	0.0	6	DNA replication, restriction/modification and repair	12.5	<i>dnaE</i>
				DNA recombination, packaging and segregation	12.5	<i>yrvE</i>
				Metabolism of amino acids and related molecules	25.0	<i>proB argH</i>
				Metabolism of nucleotides and nucleic acids	12.5	<i>yhaM</i>
				Protein synthesis and modification	25.0	<i>thrZ argS</i>
				Protein secretion	12.5	<i>csaA</i>
59	8	12.5	1	Mobility and chemotaxis	62.5	<i>yvyG yvyF flgK flgM flgL</i>
				Unknown	37.5	<i>ybdO ylbB yjcQ</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
60	5	0.0	3	Cell wall	20.0	<i>tagD</i>
				Metabolism of carbohydrates and related molecules	40.0	<i>gntZ yqeC</i>
				Metabolism of lipids	40.0	<i>ykwC yfjR</i>
61	7	0.0	4	Sporulation, germination and cell division	14.2	<i>spoVAE</i>
				Detoxification	14.2	<i>ydhE</i>
				Unknown	14.2	<i>ymfA</i>
				Protein synthesis and modification	14.2	<i>gcp</i>
				Metabolism of coenzymes and prosthetic groups	42.8	<i>ribR ribA ribC</i>
62	4	0.0	1	Metabolism of amino acids and related molecules	100.0	<i>hisI hisG hutU hisD</i>
63	4	0.0	2	Transformation/competence	25.0	<i>comER</i>
				Metabolism of amino acids and related molecules	75.0	<i>proI proH proG</i>
64	4	0.0	2	Metabolism of amino acids and related molecules	25.0	<i>mmsA</i>
				Unassigned	25.0	<i>iolF</i>
				Metabolism of carbohydrates and related molecules	50.0	<i>iolC idh</i>
65	4	75.0	1	Cell wall	100.0	<i>dacB dacF dacA dacC</i>
66	4	0.0	1	Metabolism of amino acids and related molecules	100.0	<i>leuA leuD leuC ilvH</i>
67	4	50.0	0	Unassigned	100.0	<i>licA licB ydhM ydhN</i>
68	4	0.0	1	DNA replication, restriction/modification and repair	100.0	<i>priA dnaB dnaD dnaG</i>
69	4	25.0	2	Transport/binding proteins and lipoproteins	75.0	<i>rbsC rbsD rbsB</i>
				RNA synthesis and modification	25.0	<i>rbsR</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
70	4	0.0	3	Transport/binding proteins and lipoproteins	25.0	<i>ytrF</i>
				RNA synthesis and modification	25.0	<i>ytrA</i>
				Metabolism of carbohydrates and related molecules	50.0	<i>ytrC ytrD</i>
71	4	0.0	0	Unassigned	100.0	<i>levD levG levF levE</i>
72	4	0.0	0	Unknown	75.0	<i>yanB yxbA yxbB</i>
				Unassigned	25.0	<i>yxaM</i>
73	4	0.0	1	RNA synthesis and modification	100.0	<i>ylaC sigM sigV sigX</i>
74	4	0.0	1	DNA recombination, packaging and segregation	100.0	<i>addA recF recN addB</i>
75	4	0.0	2	Unknown	50.0	<i>ykuN ytcD</i>
				RNA synthesis and modification	25.0	<i>hxlR</i>
				Metabolism of lipids	25.0	<i>ydeP</i>
76	4	0.0	2	Metabolism of nucleotides and nucleic acids	25.0	<i>pyrH</i>
				Protein synthesis and modification	75.0	<i>rpsB frr tsf</i>
77	4	0.0	1	Detoxification	75.0	<i>yceE yceF yceC</i>
				Unknown	25.0	<i>yceG</i>
78	4	0.0	0	Unknown	100.0	<i>yojI ypnP yisQ yoeA</i>
79	4	0.0	1	Adaptation to atypical conditions	25.0	<i>pspA</i>
				Unknown	75.0	<i>ydjI ydjG ydjH</i>
80	4	0.0	1	Membrane bioenergetics (electron transport chain and ATP synthase)	100.0	<i>narG narJ narI narH</i>
81	4	0.0	0	Unknown	100.0	<i>ywqK ywqH ywqI ywqJ</i>
82	4	0.0	1	Sporulation, germination and cell division	100.0	<i>gerPB gerPF gerPD gerPA</i>
83	4	50.0	1	Protein secretion	100.0	<i>sipS sipV sipU sipT</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
84	5	20.0	3	Sporulation, germination and cell division	20.0	<i>spsJ</i>
				Cell wall	20.0	<i>yfnG</i>
				Metabolism of carbohydrates and related molecules	60.0	<i>xynB xynD galE</i>
85	4	0.0	1	RNA synthesis and modification	100.0	<i>alsR yxjO yrdQ yoaU</i>
86	4	0.0	1	RNA synthesis and modification	100.0	<i>arsR ydeT yozA yvbA</i>
87	4	0.0	2	Metabolism of sulfur	75.0	<i>yitA yisZ yitB</i>
				Metabolism of amino acids and related molecules	25.0	<i>cysH</i>
88	4	25.0	1	Unknown	75.0	<i>yscB yviF yjcP</i>
				Metabolism of carbohydrates and related molecules	25.0	<i>yfmT</i>
89	4	25.0	3	Adaptation to atypical conditions	25.0	<i>degR</i>
				Transformation/competence	25.0	<i>med</i>
				RNA synthesis and modification	50.0	<i>sinR comK</i>
90	4	0.0	2	Sporulation, germination and cell division	50.0	<i>yacA ytgP</i>
				RNA synthesis and modification	25.0	<i>trmU</i>
				Unknown	25.0	<i>yhcT</i>
91	4	0.0	1	Transport/binding proteins and lipoproteins	50.0	<i>yeaB ydbO</i>
				Unassigned	50.0	<i>czcD ydfM</i>
92	5	0.0	1	Detoxification	20.0	<i>mmr</i>
				Unassigned	80.0	<i>bmr yusP yitG yceJ</i>
93	11	0.0	1	RNA synthesis and modification	100.0	<i>ywqM yofA ytlI ywbI yybE ycgK ywfK ccpC citR yvbU yclA</i>
94	3	0.0	1	Metabolism of carbohydrates and related molecules	100.0	<i>iolH iolE iolI</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
95	4	25.0	2	Membrane bioenergetics (electron transport chain and ATP synthase)	25.0	<i>yfmJ</i>
				Unassigned	25.0	<i>ydbA</i>
				Metabolism of carbohydrates and related molecules	50.0	<i>ispE ybbD</i>
96	3	0.0	1	Metabolism of coenzymes and prosthetic groups	100.0	<i>menH menD menB</i>
97	3	0.0	1	Metabolism of carbohydrates and related molecules	100.0	<i>licH malA lplD</i>
98	3	0.0	1	RNA synthesis and modification	100.0	<i>iolR yulB fruR</i>
99	3	0.0	1	Antibiotic production	100.0	<i>pksI pksH pksF</i>
100	3	0.0	1	Metabolism of amino acids and related molecules	100.0	<i>hutH hisF hisJ</i>
101	3	0.0	1	Metabolism of amino acids and related molecules	100.0	<i>asnO asnB asnH</i>
102	3	0.0	1	Membrane bioenergetics (electron transport chain and ATP synthase)	100.0	<i>atpG atpA atpI</i>
103	3	66.6	1	Cell wall	100.0	<i>ponA pbpF pbpC</i>
104	3	0.0	1	Metabolism of carbohydrates and related molecules	100.0	<i>yisS iolD iolB</i>
105	3	33.3	0	Unassigned	100.0	<i>yxeM ytmL yqiY</i>
106	3	0.0	1	Mobility and chemotaxis	100.0	<i>fliE flgC fliF</i>
107	3	0.0	2	Membrane bioenergetics (electron transport chain and ATP synthase)	66.6	<i>yyaE yoaE</i>
				Metabolism of amino acids and related molecules	33.3	<i>nasC</i>
108	3	0.0	2	Sporulation, germination and cell division	33.3	<i>spo0E</i>
				RNA synthesis and modification	66.6	<i>abrB sigA</i>
109	3	0.0	0	Unassigned	100.0	<i>pstC pstS pstA</i>
110	3	33.3	0	Unknown	100.0	<i>yycH yycI yycJ</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
111	3	0.0	1	Cell wall	100.0	<i>tagF tagA tagB</i>
112	3	0.0	1	Transport/binding proteins and lipoproteins	33.3	<i>aapA</i>
				Unassigned	66.6	<i>ybxG ybeC</i>
113	3	0.0	1	RNA synthesis and modification	100.0	<i>mta yyaN yraB</i>
114	3	0.0	1	Adaptation to atypical conditions	100.0	<i>cspC cspD cspB</i>
115	3	0.0	1	Phage-related functions	100.0	<i>zhkB xepA zhIA</i>
116	3	33.3	0	Unassigned	100.0	<i>yurN yurM yurO</i>
117	3	0.0	2	Metabolism of amino acids and related molecules	66.6	<i>yrrO yrrN</i>
				Metabolism of coenzymes and prosthetic groups	33.3	<i>yrrM</i>
118	3	0.0	0	Unknown	100.0	<i>ylaA ylaB ylaD</i>
119	3	0.0	1	Protein synthesis and modification	100.0	<i>rpmGB rplI rpmA</i>
120	3	0.0	2	Cell wall	66.6	<i>cwlH cwlA</i>
				Phage-related functions	33.3	<i>xlyB</i>
121	3	0.0	2	DNA replication, restriction/modification and repair	33.3	<i>adaA</i>
				RNA synthesis and modification	66.6	<i>yfiF ytdP</i>
122	3	0.0	2	Transport/binding proteins and lipoproteins	66.6	<i>feuC feuB</i>
				RNA synthesis and modification	33.3	<i>ybbB</i>
123	3	0.0	1	Unknown	33.3	<i>ylqD</i>
				RNA synthesis and modification	66.6	<i>rimM trmD</i>
124	3	0.0	1	RNA synthesis and modification	66.6	<i>ycnC yerO</i>
				Unknown	33.3	<i>yuzN</i>
125	3	0.0	1	DNA replication, restriction/modification and repair	100.0	<i>ykoU ligB ligA</i>
126	3	0.0	1	DNA replication, restriction/modification and repair	33.3	<i>ydiS</i>
				Unknown	66.6	<i>ydjA ydiR</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
127	3	33.3	2	Sporulation, germination and cell division	33.3	<i>tasA</i>
				Unknown	33.3	<i>yqzM</i>
				Protein secretion	33.3	<i>sipW</i>
128	3	0.0	0	Unknown	100.0	<i>yotE yotF yotD</i>
129	3	0.0	1	Transport/binding proteins and lipoproteins	33.3	<i>yhaU</i>
				Unknown	66.6	<i>yhaT yhaS</i>
130	3	0.0	0	Unknown	100.0	<i>yvcA yvcB yvzA</i>
131	3	33.3	1	Adaptation to atypical conditions	100.0	<i>ywqC ywqD ywqE</i>
132	3	0.0	1	Transformation/competence	100.0	<i>comFB comFA comFC</i>
133	3	0.0	0	Unknown	66.6	<i>yurX yurU</i>
				Unassigned	33.3	<i>yurY</i>
134	3	33.3	1	Sporulation, germination and cell division	33.3	<i>divIC</i>
				Unknown	66.6	<i>yabQ yabP</i>
135	3	0.0	1	Metabolism of carbohydrates and related molecules	100.0	<i>malS mleA ytsJ</i>
136	3	0.0	1	Metabolism of amino acids and related molecules	33.3	<i>metE</i>
				Unknown	66.6	<i>yxjH yxjG</i>
137	3	0.0	1	DNA replication, restriction/modification and repair	33.3	<i>holB</i>
				Unknown	66.6	<i>yabA yazA</i>
138	3	0.0	0	Unknown	100.0	<i>yueB yukC yukD</i>
139	3	33.3	1	Metabolism of sulfur	33.3	<i>ssuD</i>
				Unassigned	66.6	<i>ssuC ssuA</i>
140	3	33.3	1	Unknown	66.6	<i>yneA ynzC</i>
				Transposon and IS	33.3	<i>yneB</i>

Cluster	# nodes	% secreted	# functions	Function	% function	Genes
141	3	33.3	0	Unknown	66.6	<i>yufQ yufP</i>
				Unassigned	33.3	<i>yufN</i>
142	3	0.0	1	RNA synthesis and modification	100.0	<i>yurK yvoA ymfC</i>

Appendix G

Cross-species PFINs

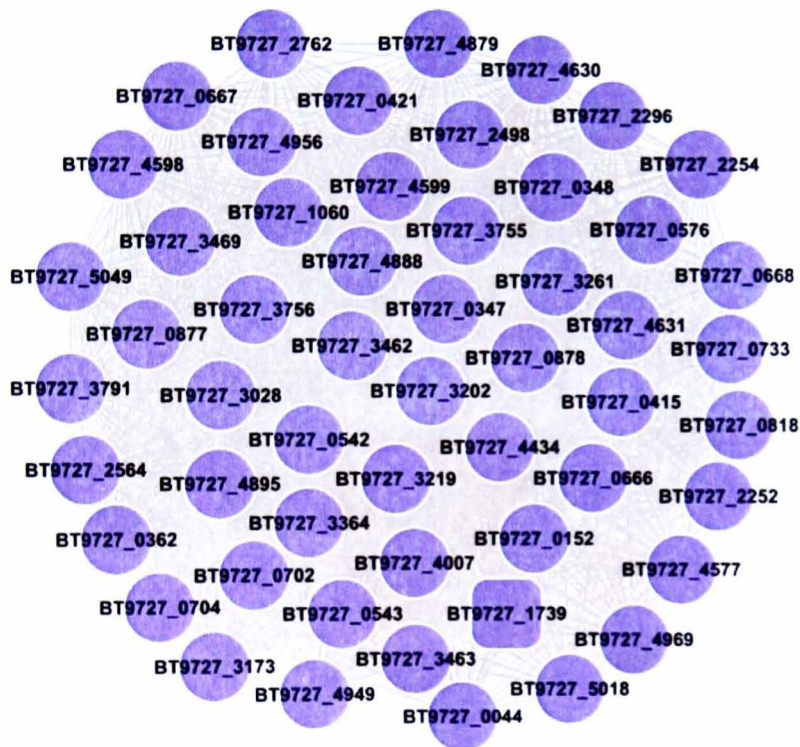


Figure G.1: The cross-species PFIN of the pathogenic protein family E.

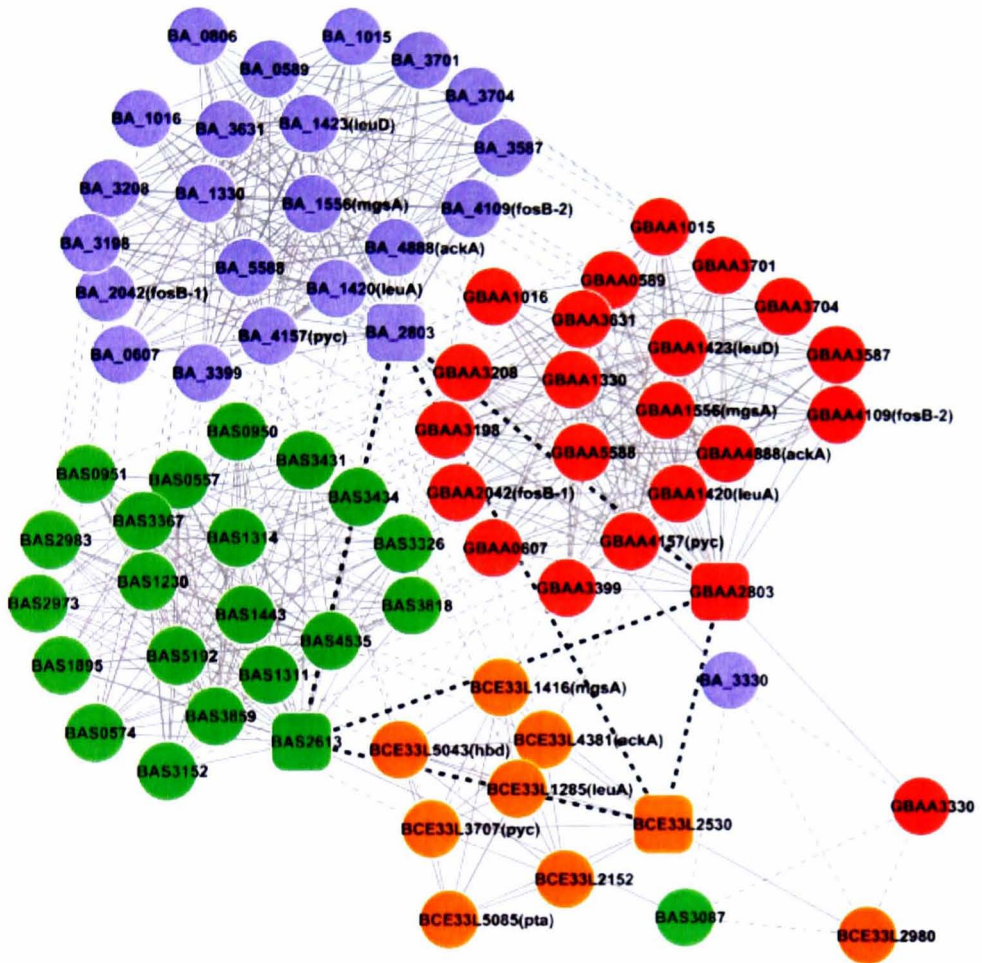


Figure G.2: The cross-species PFIN of the pathogenic protein family A.

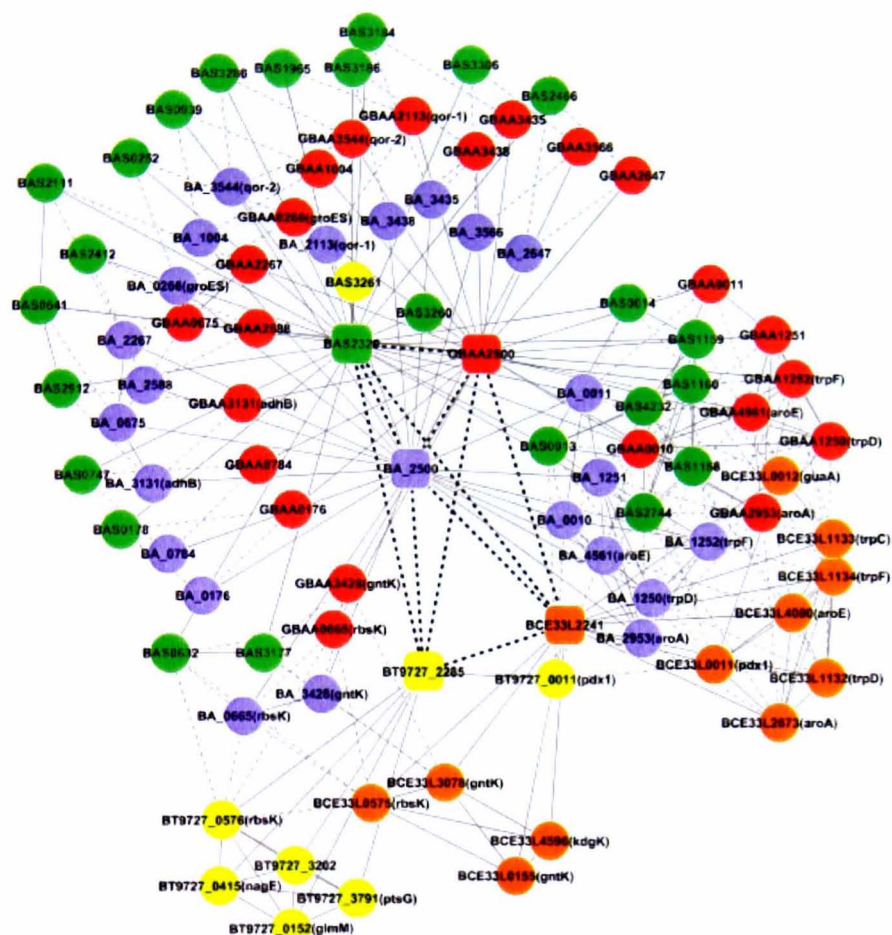


Figure G.3: The cross-species PFIN of the pathogenic protein family L.

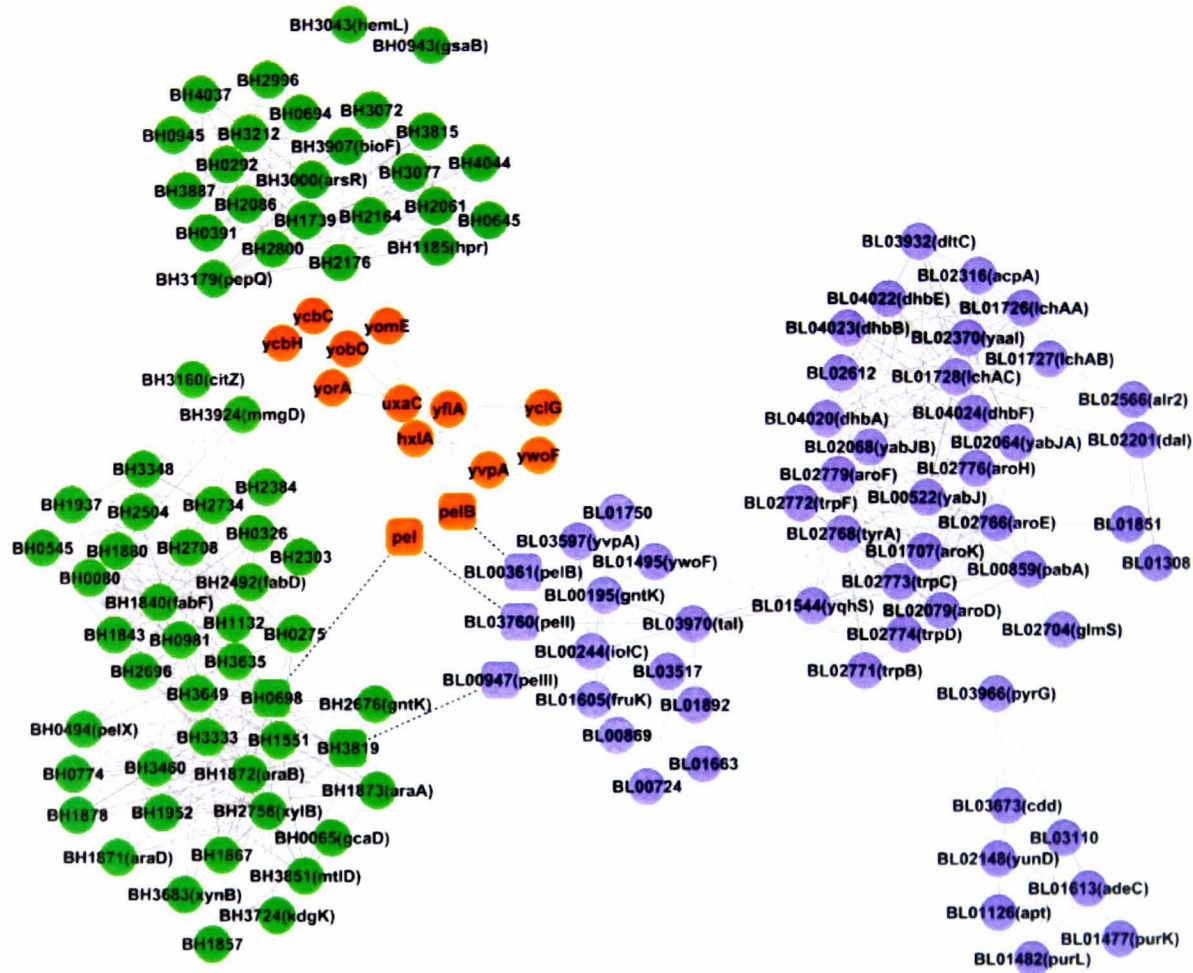


Figure G.4: The cross-species PFIN of the non-pathogenic *pel* protein family.